The Thermodynamics and Kinetics of Protein Folding: A Lattice Model Analysis of Multiple Pathways with Intermediates

Aaron R. Dinner[†,‡] **and Martin Karplus***[,†,‡,§]

*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street,
Cambridge, Massachusetts 02138, Committee on Higher Degrees in Biophysics, Harvard University,
Cambridge, Massachusetts 02138, and Laboratoire de Chimie Biophysique, Institut le Bel,
Université Louis Pasteur, 4 Rue Blaise Pascal, 67000 Strasbourg, France*

Received: March 10, 1999; In Final Form: May 5, 1999

The kinetics and thermodynamics of folding of a representative sequence of a 125-residue protein model subject to Monte Carlo dynamics on a simple cubic lattice were investigated. The diverse trajectories that lead to the native state can be classified into a relatively small number of average pathways: a "fast track" in which the chain forms a stable core that folds directly to the native state and several "slow tracks" in which particular contacts form before the core is complete and direct the chain to a misfolded intermediate. Rearrangement from the intermediates to the native state is slow because it requires breaking stable contacts, which involve primarily surface residues. The transition state for folding is identified by activated dynamics simulations and consists of a reduced version of the core in the absence of other (native and nonnative) contacts which slow folding. Each track involves an ensemble of structures that can be characterized by two progress coordinates for the reaction. These coordinates are based on a comparison of folding and nonfolding trajectories: one coordinate monitors the formation of the core and the other monitors whether the chain is trapped in a long-lived intermediate. From Monte Carlo simulations, we obtain an estimate for the density of states and calculate equilibrium averages, including the free energy, energy, and entropy, as functions of the two coordinates. The thermodynamics are in good agreement with the observed kinetics; the transition states correspond to plateaus or barriers in the free energy while the intermediates are energetically stabilized local free energy minima. The complexities of the folding mechanism bear a striking similarity to those observed experimentally for lysozyme, a well-studied protein of comparable size.

1. Introduction

A full description of a reaction requires a mapping of all the intermediates that lead from the reactants to the products and the transition states that connect them. Although explicit consideration of each possible system configuration is feasible for the simplest reactions, such as H$-$H$_2$ exchange,[1] it is not for processes of greater complexity. Instead, it is necessary to group states by projecting the configuration space onto a small number of (reaction) coordinates that characterize the system. If suitable reaction coordinates can be found, the free energy profile for each step in the reaction can be used to determine the kinetics with activated dynamics techniques.[2,3] This has been done for reactions in the gas phase,[4,5] for reactions in solution,[6,7] and for reactions in proteins.[8,9]

However, reaction coordinates are often difficult to determine. Even for a reaction as simple as the flip of an aromatic ring in the bovine pancreatic trypsin inhibitor (BPTI),[8] it was found that the "obvious" coordinate (the dihedral angle of the ring) is inadequate. Analysis of reactive and non-reactive trajectories revealed that it is necessary to construct a more complex coordinate that accounts for the interaction of the side chain with the main chain.[8] For reactions such as the ring flip, which are relatively localized, automatic methods for determining the reaction path have been developed. However, most automatic methods consider only the potential energy (for example, see ref 10), so that they are not suitable for processes with significant entropic effects. Methods that do account correctly for the free energy either rely on a knowledge of the minimum energy path and a reasonable guess for the reaction coordinate[11] or they are computationally very demanding because they essentially require simultaneous consideration of all time points.[12]

One reaction which is particularly difficult to describe is the folding of a protein from its denatured (coil) state to its unique, three-dimensional native structure.[13] Both experiments and all-atom simulations are being used to characterize the reactant (the denatured state[14]), the intermediates,[15] and the product (the biologically functional native state[3]). Because the entire macromolecular system participates in the reaction and entropy plays a substantial role, it is not feasible to employ the automatic methods mentioned above. Finding a reduced set of coordinates that is adequate for describing the reaction is an essential (and challenging) element of the analysis. Moreover, even with such a set of coordinates, the definition of the transition state region(s) is not straightforward.

These difficulties have led to the introduction of simplified protein models, for which one can obtain many folding (reactive) and nonfolding (nonreactive) trajectories that can be used to determine the reaction path(s). In most such simplified model studies, the chain is restricted to a lattice and the residues are represented by single points (beads) that interact with a nearest-

* Correspondence: Martin Karplus. E-mail: marci@tammy.harvard.edu. Phone: (617) 495-4108. Fax: (617) 496-3204.

[†] Department of Chemistry and Chemical Biology.

[‡] Committee on Higher Degrees in Biophysics.

[§] Laboratoire de Chimie Biophysique.

10.1021/jp990851x CCC: $18.00
Published on Web 08/31/1999

neighbor potential of mean force (for reviews, see refs 16−19). Šali and co-workers studied one such system: a 27-residue chain subject to Monte Carlo (MC) dynamics on a simple cubic lattice.[20] At low temperatures, where the native state is stable, folding proceeds by a fast ($\sim 10^4$ MC steps) collapse to a semicompact random globule, followed by a slow ($\sim 10^7$ MC steps) nondirected search through the ($\sim 10^{10}$) semicompact structures for one of the ($\sim 10^3$) transition states that lead rapidly ($\sim 10^5$ MC steps) to the native conformation. By using MC to sample the equilibrium population at a single (elevated) temperature, these authors obtained the density of states as a function of the number of native contacts (Q_0) and calculated the free energy as a function of that variable. The free energy exhibits two minima whose relative importance depends on the temperature: an entropic one at $Q_0 \approx 6$ (out of 28), corresponding to the denatured state, and an energetic one at $Q_0 = 28$, corresponding to the native state. The free energy barrier separating the minima ($Q_0 \approx 23$) coincides with the bottleneck in independent kinetic trials, at which point the probability of reaction increases from essentially zero to near unity[20,21] (A. Šali and M. Karplus, unpublished). The agreement between the thermodynamically and kinetically determined transition states indicates that Q_0 is a suitable variable for describing the folding reaction in this case (i.e., for the 27-mer with interactions drawn randomly from a Gaussian distribution).

The 27-mer folding mechanism appears to be limited to small proteins. Since the folding time is dominated by the random search, it scales with the ratio of compact configurations to transition states. This ratio is expected to increase exponentially with chain length and yield unrealistically long folding times for chains of more than about 80 residues.[20] To determine how longer chains might fold, we studied 125-mers with native states that are $5 \times 5 \times 5$ cubes with significant secondary structure. The study of a few such sequences[22,23] showed that they fold by a mechanism in which the chain quickly ($<10^6$ MC steps) collapses to a disordered globule and then makes a relatively slow search ($\sim 10^7 - 10^8$ MC steps) through the compact states for a specific set of about 30 contacts (nonbonded spatial nearest neighbors) which make up a stable, cooperative core that results in a rapid accumulation of native structure. Once the core is formed, both direct folding to the native state and misfolding to long-lived intermediates occur. Molecules which form an intermediate complete folding by rearrangement and condensation of primarily surface residues; this is slow ($\sim 10^7 - 10^9$ MC steps) because it requires disruption of stable contacts. The generality of these results was confirmed by statistical analysis of a 200 sequence database which showed that folding is promoted by increases in the overall stability, cooperativity, and kinetic accessibility of the native state and is hindered by overstabilization of the contacts between surface residues.[22,24]

We report here the detailed analysis of a representative 125-mer sequence that folds relatively slowly and often becomes trapped in an intermediate. Thus, the features of the free energy surface that lead to the two slow steps (core formation and surface rearrangement) should be pronounced. It is evident that a single progress variable cannot be sufficient to distinguish trajectories that fold directly to the native state from those that go through an intermediate. This idea is consistent with studies of related systems[19,25] that have shown that Q_0 does not provide an adequate description of the reaction when configurations with the same Q_0 exhibit markedly different kinetics of interconversion to the native state. It has been suggested that the probability that each structure will fold before unfolding (p, the transmission coefficient) be used as the variable that monitors the progress of the reaction.[18,25] Doing so solves the problem with using Q_0 because, by definition, all the structures at a given p have the same likelihood of completing the reaction. Structural features of the transition states and the intermediates can then be obtained by averaging over configurations in a particular range of p. However, p is only a formal solution to the problem of finding a reduced set of coordinates because determination of p requires simulation of the reaction dynamics starting with each configuration of the system, so that the complexity of the calculation is not really reduced (unless additional approximations are made).

In the present analysis, we proceeded in a manner analogous to that used in the study of aromatic ring flips in BPTI[8] and chose our reduced set of coordinates to reflect structural differences between reactive (folding) and nonreactive (nonfolding) trajectories. Two (or more) coordinates are required to separate trajectories that form a long-lived intermediate (slow or nonfolding) from those that do not (fast folding). We took the first coordinate to be a subset of native contacts that is present with high probability immediately prior (in Q_0) to core formation in folding trajectories and the second coordinate to be a subset of native contacts that appears with higher probability in folding trajectories than that in non-folding trajectories at values of Q_0 corresponding to the misfolded intermediates. Although the trajectories that are sampled in reaching the native state are quite diverse, they can be classified into several average pathways using these two coordinates: a "fast track" in which the chain forms a stable core that folds directly to the native conformation and several "slow tracks" in which, prior to completing formation of the core, the chain makes particular noncore (native and nonnative) contacts that direct the chain to one of several misfolded intermediates that partially unfold to reach the native state. The results have a striking similarity to the experimental folding behavior of lysozyme,[13,26] a well-studied protein which is comparable in length to the 125-mer (129 residues).

2. Choice of Coordinates

To obtain an adequate reduced description of the free energy, energy, and entropy surfaces for the folding reaction, it is necessary to find suitable coordinates, as pointed out in the Introduction. Reaction coordinates for complex systems are composed of the smallest subset of the degrees of freedom that is sufficient to describe the slow steps in the conversion of reactants to products. The existence of such "slow" coordinates is an essential element in the description of the reaction in terms of transition state theory and its generalized analogues.[27] Included in this description is an implicit canonical average over all other degrees of freedom. This leads to surfaces that depend on temperature even though the interaction terms in the present model do not. For most complex systems, including protein folding, it is expected that only a few degrees of freedom will be adequate to delineate the essential features of the reaction. For example, a linear combination of two dihedral angles was found to be sufficient to express the potential of mean force for the flip of an aromatic ring in BPTI.[8] In that case, the selected coordinates were determined by a trial-and-error approach based on analyzing reactive and nonreactive trajectories.

A corresponding approach can be used to find suitable coordinates for the folding of a lattice model. For the 27-mer, the reaction is simple; even though there are many possible trajectories that lead to the native state, there are no long-lived "off-pathway" intermediates. As a result, a single coordinate (the number of native contacts, Q_0) is satisfactory, as discussed

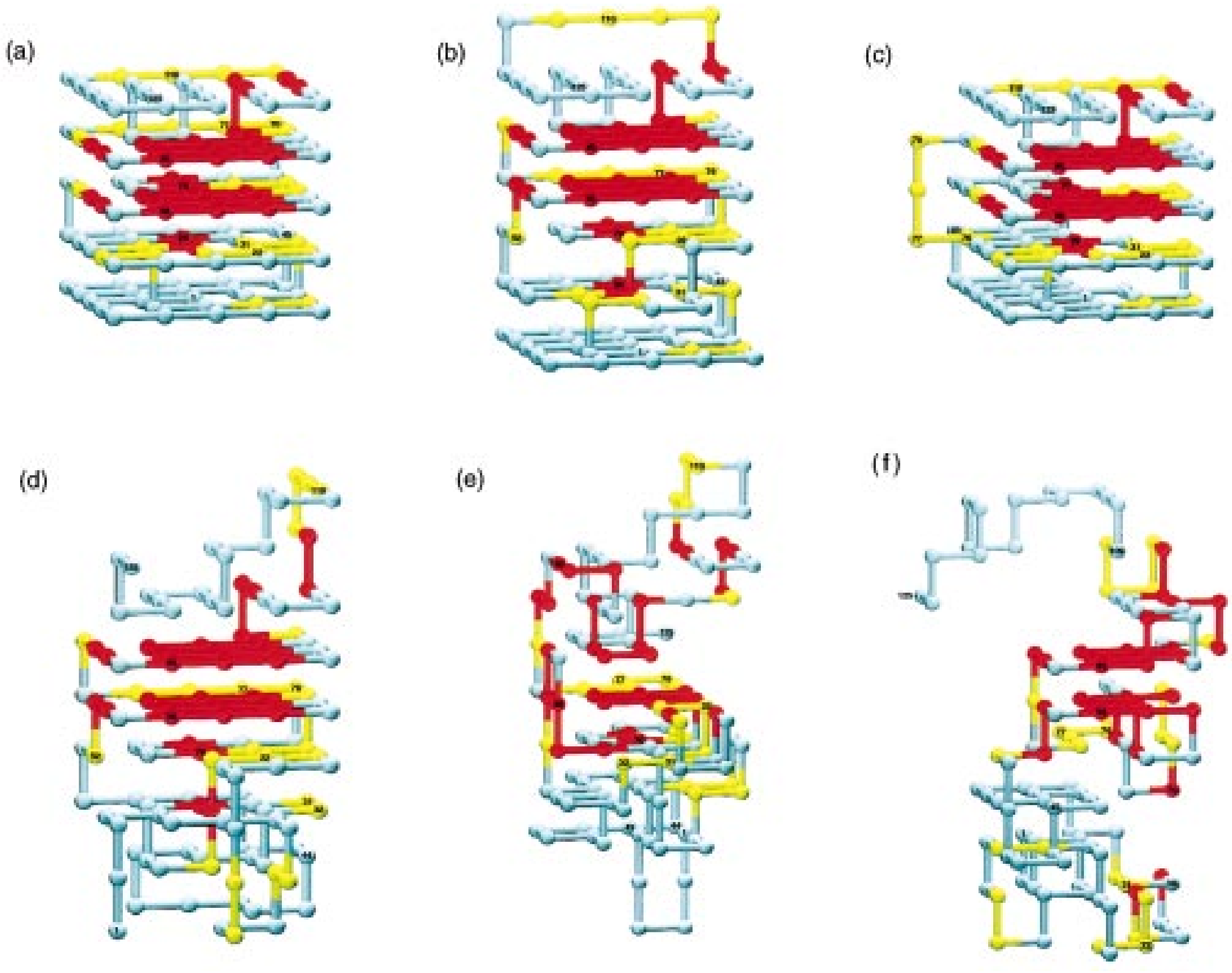

Figure 1. Structures sampled during folding. Residues in the core are red. Residues contributing to Q_s, contacts are yellow. All others are blue. (a) Native conformation: $E = -224.157$, $Q_0 = 176$, $Q_c = 34$, $Q_s = 15$, $C = 176$, $C_{as} = 81$, and $C_{ps} = 76$. (b) Intermediate I_1: $E = -199.842$, $Q_0 = 147$, $Q_c = 33$, $Q_s = 0$, $C = 162$, $C_{as} = 84$, and $C_{ps} = 57$. (c) Intermediate I_2: $E = -200.382$, $Q_0 = 150$, $Q_c = 34$, $Q_s = 11$, $C = 164$, $C_{as} = 76$, and $C_{ps} = 68$. (d) The most unfolded structure from the trial that unfolded the least during interconversion from I_1 to the native state at $T = 1.0$: $E = -140.218$, $Q_0 = 89$, $Q_c = 32$ and $Q_s = 0$, $C = 129$, $C_{as} = 59$, and $C_{ps} = 37$. (e) A typical member of the N_1 transition state ensemble: $E = -94.162$, $Q_0 = 41$, $Q_c = 21$, $Q_s = 0$, $C = 103$, $C_{as} = 27$, and $C_{ps} = 27$. (f) A typical member of the I_1 transition state ensemble: $E = -113.757$, $Q_0 = 41$, $Q_c = 14$, $Q_s = 0$, $C = 121$, $C_{as} = 57$, and $C_{ps} = 20$. The structures were generated with the program VMD of the MDScope computing environment.[72]

in the Introduction. For the 125-mer, the reaction is more complex; there are multiple paths, some of which lead directly to the native state and others that involve intermediates. To describe such behavior, additional (structural) coordinates must be introduced. In the present study, it is shown that two coordinates are sufficient to describe the essential aspects of most trajectories. In contrast to the BPTI example,[8] these coordinates cannot be reduced to a single linear combination that describes the entire reaction. Rather, there exists a two-dimensional surface on which certain combinations of the coordinates tend to lead to one reaction pathway and different combinations to others. If there was a perfect one-to-one correspondence between each point on this two-dimensional surface and the reaction pathways, it would be possible to find a transformation which reduced the surface to a series of truly one-dimensional paths (and the reaction coordinates would measure the distance along those paths). However, due to both the inherent uncertainty of the stochastic processes involved and the practical limitations of our two-dimensional structural analysis, such a mapping does not exist (i.e., trajectories contributing to different pathways sometimes overlap). To reflect this complexity, we refer to the pair of coordinates as "progress variables" rather than "reaction coordinates". This fundamental complexity in theoretical approaches to understanding a protein folding reaction is illustrated by the present analysis of the 125-mer.

Because of the length of the required simulations, we present calculations for only one sequence (details of the model and the simulations are presented in Supporting Information). This sequence was chosen because, as mentioned above and detailed below, both slow steps of the folding mechanism are pronounced, but it is still capable of folding in most trials (of length $\sim 10^8$ MC steps). Its native structure (Figure 1a) is composed entirely of sheets, with 76 contacts in parallel sheets and 81 contacts in antiparallel sheets (30 of which are in turns) (see ref 28 for definitions and Figure 3S, Supporting Information, above diagonal). The residues are grouped sequentially in the structure; the first 50 residues fall in the bottom two horizontal planes of Figure 1, residues 51−75 make up the middle plane, residues 76−98 account for most of the second plane from the top, and residues 99−125, with the exceptions of 121 and 122, make up the top plane. Most of the antiparallel sheet contacts

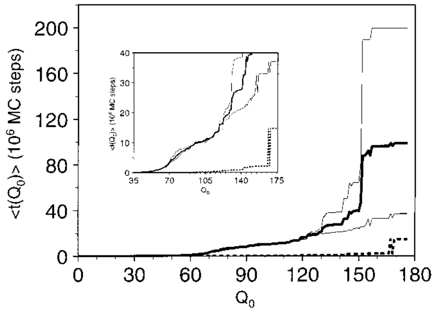

Figure 2. Average first step [$\langle t(Q_0)\rangle$] at which a given contact overlap with the native state (Q_0) is reached. Solid lines are for trials beginning in random configurations; the overall average (bold) is based on 50 trials, of which 31 reach the native state (lower thin) and 19 do not (upper thin). The dashed line is for trials in which the core contacts are introduced; the other contacts are randomized at high temperature ($T = 10\ 000$), and folding is subsequently allowed to proceed without any constraints at the normal folding temperature ($T = 0.8$); only one line is shown because all 20 such trials fold. For clarity, error bars are not shown but are typically less than 20% of the average at that Q_0. Since a chain may skip some Q_0, those $t(Q_0)$ are taken to be the $t(Q_0)$ of the next sampled Q_0; $t(Q_0)$ for Q_0 higher than the maximum reached in that trial are assigned the maximum number of allowed steps (200×10^6 MC steps).

are in hairpins that lie within the planes, while most of the parallel sheet contacts serve to connect the planes.

Since the appropriate choice of progress variables depends on the mechanism of the reaction, it is necessary to use simulations to determine the specific folding behavior of this sequence. We performed 50 MC trials at the same temperature as in earlier studies ($T = 0.8$),[22,24] each of which began in a random configuration and proceeded for 200×10^6 MC steps or until the native state was reached (first passage time). Thirty-one of the trials reached the native state, and 19 did not. In Figure 2, we show the mean first-passage time of each Q_0 [$\langle t(Q_0)\rangle$]. Plateaus indicate that the chain is progressing rapidly in Q_0, while significantly sloped or near vertical segments indicate that the chain is trapped in a given Q_0 region. The bold line, which is calculated from all 50 trials, shows that there are two slow steps on the average. The first, at $60 \leq Q_0 \leq 80$, corresponds to core formation, and the second, at $Q_0 \approx 150$, corresponds to misfolding and rearrangement. Upon separating trials that reach the native state (lower thin line) from those that do not (upper thin line), we see that both sets of trials form the core but only the latter becomes trapped at $Q_0 \approx 150$. This qualitative difference in the behavior between structures with a given Q_0 (i.e., at $Q_0 \approx 150$, some interconvert rapidly to the native state, while others do not) shows that Q_0 is not an adequate progress variable for the 125-mer, in contrast to the simpler behavior found for the 27-mer. Moreover, it indicates that at least a two-dimensional representation is required to encompass the essential features of the reaction.

We chose to use structure-based progress variables and monitor subsets of the native contacts which appear more frequently in trials that reach the native state than in ones which do not. We took the first progress variable to be the number of core contacts (Q_c). The core contacts were identified in the following manner. We counted the number of times each contact (both native and nonnative) appeared in the last structures sampled for each of the Q_0 in the range $61 \leq Q_0 \leq 88$ in the 31 trials that reached the native state. Particular native contacts (about 35) were found to occur with probabilities (p_{ij}) greater than about 0.6 (the highest nonnative contact probability was 0.32). We took these contacts to be candidates for the core contacts and verified their role by fixing the residues involved, randomizing the rest of the structure by simulation for 5×10^6 MC steps at high temperature ($T = 10\ 000$), and subsequently allowing folding to proceed without any constraints at the normal folding temperature ($T = 0.8$). The set of fixed residues was modified by trial and error until a minimal set that led to consistent rapid folding was found (dashed curve in Figure 2). It contains 31 residues (Figures 1 and 3S), which make 34 contacts with each other (8 of the 34 contacts have $0.25 \leq p_{ij} \leq 0.6$ and are included because they involve residues that also make high probability contacts). Although 19 of these are in antiparallel hairpins, which have been shown to accelerate folding by restricting the random search for the core,[22,24] 11 of them fall within a single, long parallel sheet with $|i - j| = 29$ that connects residues 51−69 to residues 80−98. Such long-range contacts are more difficult to find by a random search, which slows core formation somewhat relative to faster folding sheet sequences in which the core is typically composed primarily of short-range antiparallel sheet contacts (compare Figure 2 to Figure 1b of ref 22).

To monitor whether the chain is trapped in a long-lived intermediate or not, we introduce a second progress variable (Q_s). It is a subset of native contacts that is often absent in the long-lived intermediates formed at $Q_0 \approx 150$. The contacts contributing to Q_s are identified as those which appear 40% more frequently in the range $131 \leq Q_0 \leq 150$ in the 31 trials that reach the native state than those in the 19 which do not. The cutoff of 40% was chosen by trial and error to yield a moderate number of contacts. There are 15 such contacts (Figures 1 and 3S); many, but not all, are on the surface, and one, (24−65), is in the core. It is important to note that the two sets of contacts that are used to follow the reaction Q_c and Q_s were chosen by examining what happens in the folding process rather than by an automated procedure. They are not necessarily unique but, as we show, they appear to monitor well the degrees of freedom which relax most slowly during the reaction towards the native state while averaging over all the "faster" (less interesting) degrees of freedom. Consequently, they are suitable progress variables for the reaction.

The variables Q_c and Q_s incorporate information from the kinetic trials but can be determined in a very small amount of computer time. In contrast, the transmission coefficient (p) mentioned in the Introduction is very expensive computationally to determine. Consequently, it is not feasible to use p to constrain the system during equilibrium simulations to ensure sampling of the accessible conformational space, which is necessary for the 125-mer to achieve adequate sampling of the intermediates. Specifically, we obtained an intial estimate for the density of states as a function of E, Q_c, and Q_s by adding a harmonic (umbrella) potential to constrain the chain to particular values of structural variables and then refined this estimate in a series of multicanonical trials (see Supporting Information for descriptions of these methods and details of the simulations). Prior to the calculation presented, we tried sampling by the J-walking technique[29] without constraints at several temperatures and by standard metropolis Monte Carlo[30] with harmonic constraints in the total number of contacts (C) alone, in Q_0 alone, in Q_c alone, in both Q_0 and C, and in both Q_0 and Q_c. All of these simulations began in random configurations and were cooled slowly as described for the umbrella sampling trials in the Supporting Information. Although the trials appeared to be at equilibrium, there was little evidence of the intermediates observed in the kinetic trials. After realizing the need to constrain the system in Q_s, we went back and analyzed the previous simulations. In each of these cases, a quasiequilibrium had been reached which sampled only the fast-folding pathway. The intermediates tend not to be sampled because, as we will show, they are high free energy local minima, from which the chain readily escapes when the system is cooled sufficiently slowly. The existence of such pitfalls in this relatively simple model suggests that some care is required in interpreting free energy surfaces of reactions for which the kinetics have not been studied.[31,32]

3. Kinetic Results

To analyze the progress of the reaction, we map the folding mechanism onto the Q_cQ_s-plane. We base much of our analysis on the last structure sampled at each Q_0 in each trial. In trials in which the chain makes progress towards the native state (as measured by Q_0), the last structures sampled at each Q_0 are those that lead directly to higher Q_0, and in trials in which the chain becomes trapped, the last structures sampled at each Q_0 (which are different from the former set of structures) are those that cannot interconvert to higher Q_0 without at least partial unfolding. We focus on structures saved at regular intervals in Q_0, rather than ones saved at regular intervals in time (MC steps), because the chain can spend a very long time at a particular bottleneck in one trial but only a relatively short time at the same point in another trial due to the stochastic nature of the simulations. However, the differences observed for folding and nonfolding trials (Figure 2) do not stem from the fact that the time required for a given structure to interconvert to the native state varies from trial to trial but instead from the fact that the folding and nonfolding trials sample separate sets of structures that differ in their average ability to interconvert to the native state. The conclusions drawn from sampling at regular intervals in time are qualitatively similar to those drawn from sampling at regular intervals in Q_0 (data not shown).

3.1. "Fast Track". Before presenting trajectories for the trials that began with random structures, we describe the trajectories followed by the last structures at each Q_0 in the 20 trials in which the core was introduced (section 2) to clarify the role of the core and the behavior of the progress variable Q_c (Figure 3f). As stated in section 2, the core contacts are defined as a minimal set whose presence yields rapid folding (dashed line in Figure 2). These trajectories do not start at $Q_c = 34$ because, even though the heating period ends on the average at $Q_c = 34$ and $Q_s = 3.05$, the chain quickly loses a few core contacts, and the last structures at the lowest Q_0 sampled typically have $Q_c \approx 30$ (an example is shown in Figure 2Sa, Supporting Information). The contacts that are lost rapidly in these trials are not always the same; for the structures at $Q_0 = 58$, the first Q_0 sampled by all 20 trials, contact (29−70) has probability $p_{29-70} = 0.5$ and all other core contacts have higher probabilities. Most of the contacts contributing to Q_s at $Q_0 = 58$ have low probabilities, although there are a few exceptions: $p_{52-81} = 0.55$, $p_{24-65} = 0.65$, $p_{68-97} = 1$, $p_{69-98} = 1$. The last two are not geometrically required for the Q_c at which they occur, but they have very favorable energies ($B_{68-97} = -1.48$ and $B_{69-98} = -1.37$) and are easily found given the constraints of the core contacts. After the initial rearrangements, the chain accumulates native contacts rapidly and reaches $Q_0 \approx 165$ in about 2.2×10^6 MC steps on average. In the Q_cQ_s-plane (Figure 3f), the average trajectory follows an essentially straight line to the native state at ($Q_c = 34$, $Q_s = 15$). Although in 8 trials the chain folds directly to the native state, in 12, it becomes trapped at $Q_0 \approx 165$ (typically by misfolding the last 25 residues), so that the mean-first passage time of reaching the native state is substantially higher: $\langle t(Q_0 = 176)\rangle = 14.8 \times 10^6$ MC steps. These relatively short-lived traps do not produce a deviation in the path projected onto the Q_cQ_s-plane because they fall at roughly the same coordinates as the native state: (34,14) and (34,15). These 20 trials are all rapid in comparison to others (Figure 2) and correspond to a "fast track"[33] in which the chain first forms most of the core but relatively few other contacts ($Q_c \approx 30$ and $Q_s \approx 3$) and then proceeds almost directly to the native state (reflected by a rapid increase in Q_s).

We now turn to trials that begin in random configurations, and consider the 50 Monte Carlo trials of up to 200×10^6 MC steps described in section 2. In these trials, the starting configuration is typically relatively open and has between 15 and 75 contacts (out of a possible 176), of which between 0 and 12 are native. In the reduced coordinate space, it falls in the range ($-25 \le E \le -5$, $0 \le Q_c \le 4$, $Q_s = 0$). For comparison, a completely open chain would have (E, Q_c, Q_s) = (0, 0, 0), but such a conformation is almost never observed, even at higher temperatures. From the random starting conformation, the chain collapses to a compact, disordered globule with about 115 contacts within the first 1×10^6 MC steps. This collapse is about an order of magnitude slower than that described for the random 27-mer sequences[20] because the nonspecific pairwise interaction (B_0) which maximizes the overall folding rate is smaller in magnitude for the optimized 125-mer ($B_0/T \approx -0.3$ for the 125-mer compared to $B_0/T \approx -1.5$ for the 27-mer with $T = 1.33$). It is important to note that even for such a small hydrophobic term, the collapse is to a disordered globule rather than an ordered molten-globule-like state. For further discussion of the role of B_0 in this type of model, see refs 24 and 34. The disordered globule has substantial (random) overlap with the native state ($Q_0 \approx 44$) and populates the range ($-135 \le E \le -80$, $0 \le Q_c \le 16$, $0 \le Q_s \le 3$). Although it can contain a significant number of core contacts, they are not in the correct spatial arrangement (in contrast to the starting structures of the trials in which the core was introduced) and the core is disorganized.

In these trials, a substantial part of the core ($20 \le Q_c \le 30$) is formed with a mean-first passage time of about 12×10^6 MC steps. In the Q_cQ_s-plane, the trajectories traced by the last structures at each Q_0 rise more quickly in Q_c than those in Q_s (Figure 3) for both folding and nonfolding trajectories. The 50 trials can be grouped into several average pathways. The trajectories that fold are classified into two groups: those that reached the native state within 10×10^6 MC steps (7 trials, denoted N_1) and all others (24 trials, denoted N_2); the nonfolding trials are discussed below. The faster folding trajectories (Figures 3a and 4a) cluster close to the "fast track", while the slower folding trajectories exhibit more variation and deviate substantially from it (Figures 3b and 4a). In the latter, (N_2) partial formation of the core ($Q_c \approx 15$) stimulates a premature accumulation of noncore native contacts (as measured here by Q_s), which results in trapping at high Q_0. These traps (intermediates) are qualitatively similar to those described below in conjunction with the non-folding trials; they differ only in that they are shorter lived. To complete folding, the chain must partially unfold, breaking about 80 contacts on the average. Although some unfolding is observed in the fast trials (N_1) as well, it is considerably less (breaking about 40 contacts on the average) (data not shown).

3.2. Intermediates. The second slow step in the folding mechanism, when it occurs, derives from trapping in long-lived intermediate states. The coordinate Q_s monitors contacts that have substantially lower probability in the 19 kinetic trials that

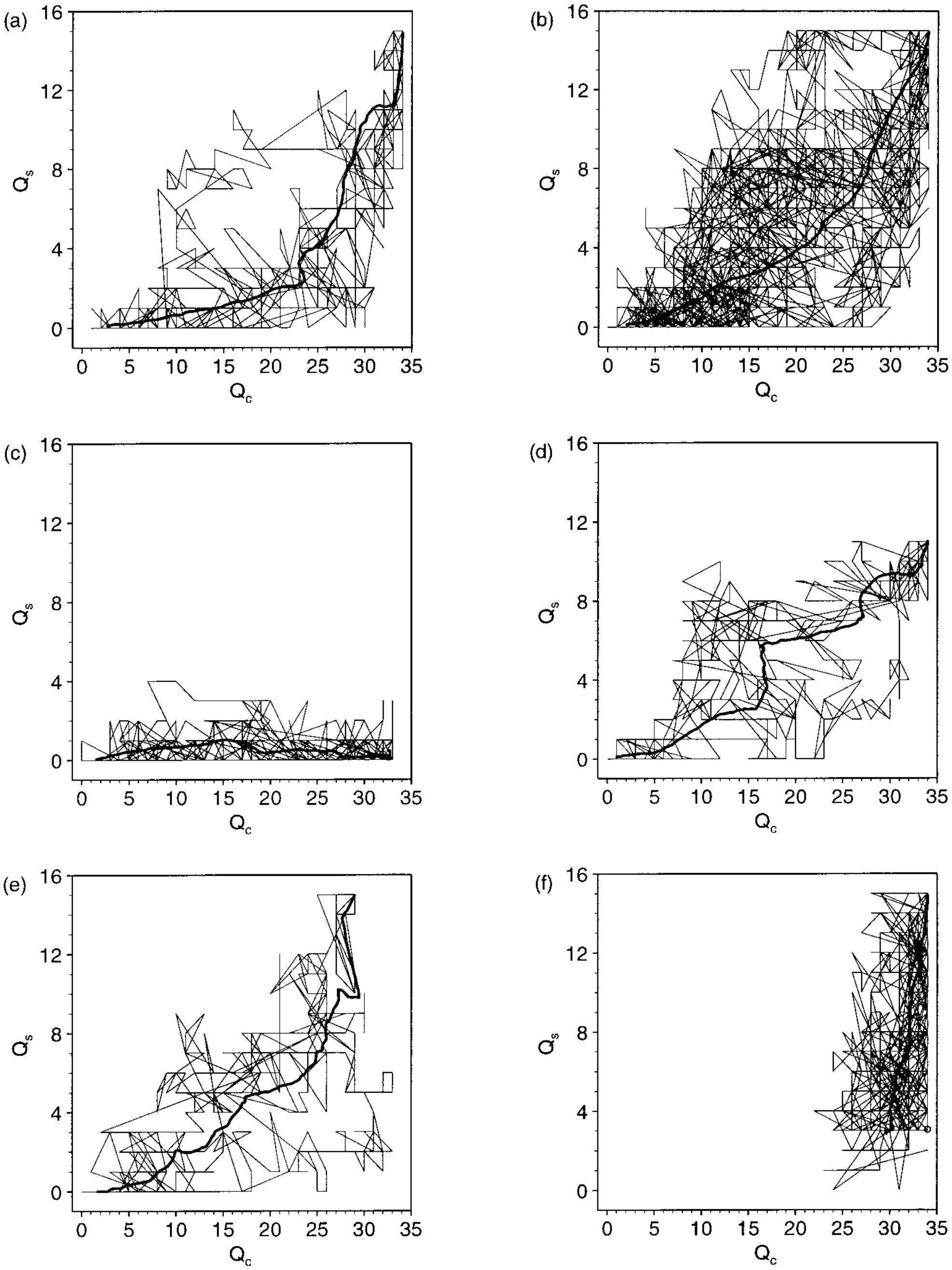

Figure 3. Trajectories at $T = 0.8$ in the Q_cQ_s-plane; they show the last structures at each Q_0 in each of 50 trials that began in random configurations and proceeded for up to 200 × 10^6 MC steps: (a) 7 trials that reach the native state within 10 × 10^6 MC steps (N_1), (b) 24 trials that reach the native state in between 10 × 10^6 and 200 × 10^6 MC steps (N_2), (c) 9 trials that reach I_1, (d) 5 trials that reach I_2, (e) 5 trials that reach intermediates other than I_1 and I_2 (I_{3+}). (f) The same for 20 trials in which the core was introduced (C) (see text); the circle indicates the average position [(34,3.04545)] immediately after the heating stage of those trials. Thin lines trace individual trajectories, and bold lines trace average trajectories; the average trajectories were smoothed by averaging over a running window of 11 Q_0 in addition to over the trials.

failed to reach the lowest energy state than in the 31 kinetic trials that succeeded (section 2). Subsequent analysis showed that these contacts have low probability in the non-folding trials because they are absent in a specific misfolded intermediate (I_1, Figure 1b) which is the endpoint of 9 of the 19 nonfolding trials (Figures 3c and 4a). For this conformation, (Q_0, Q_c, Q_s) = (147, 33, 0), and $E = -199.842$, an energy 24.315 units above the native state. The remaining 10 nonfolding trials form some or all of the contacts included in Q_s. Five of these trials folded into an intermediate with $E = -200.382$ [denoted I_2: (Q_0, Q_c, Q_s) = (150, 34, 11)] (Figures 1c, 3d, and 4a) and five folded into intermediates that were each reached only once with energies ranging from -200.753 to -184.619 (I_{3+}, where the "+" denotes the fact that I_{3+}, in contrast to I_1 and I_2, includes several different states) (Figures 3e and 4a). These intermediates are different in character from the metastable state described for another sequence,[23] in which folding from a random configuration with the kinetic move set always resulted in the metastable state, rather than the lowest energy one and interconversion did not occur within feasible simulation times (10^9 MC steps). A full description of all the intermediates observed in the present study would require additional progress variables. However, the space of reduced coordinates becomes substantially more complicated as one introduces additional dimensions, and we chose to focus on the most important intermediates, I_1 and I_2.

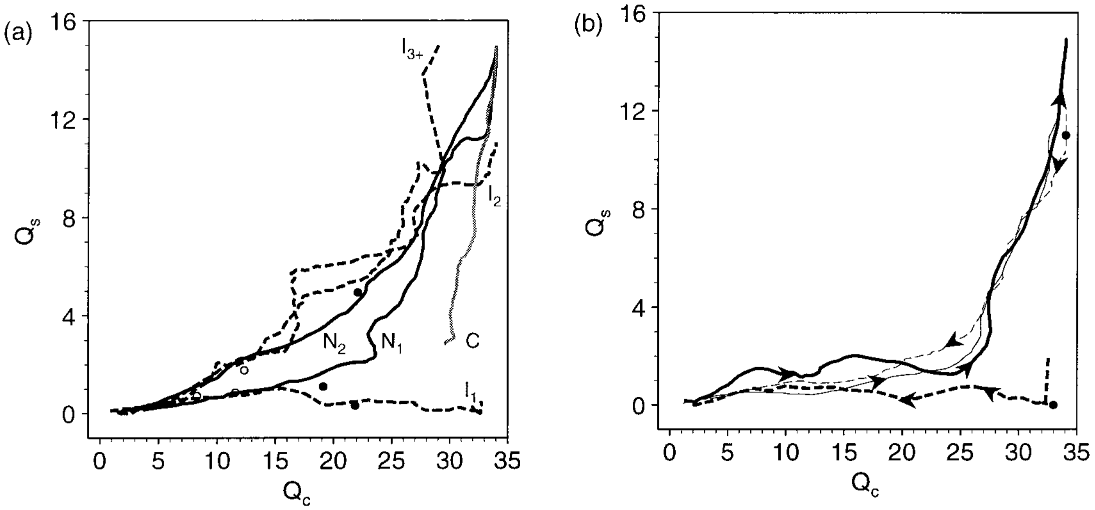

Figure 4. Average trajectories. (a) Trials shown in Figure 3. Open circles indicate the positions of the first set of transition states (T_1), and filled circles indicate the positions of the second set of transition states (T_2) (see text). (b) Trials that began in the intermediates and successfully rearranged to the native state at $T = 1.0$ (see text). (bold lines) 20 trials that began in I_1, marked by a circle at (33,0). (fine lines) Eight trials that began in I_2, marked by a circle at (34,11). Dashed lines are for the first structures at each Q_0 (unfolding from the intermediate to a denatured state), and solid lines are for the last structures at each Q_0 (folding from a denatured state to the native state); arrows are drawn to guide the eye.

The intermediates I_1 and I_2 are misfolded conformations with substantial native secondary structure (Figures 1b, 1c, and 4S (Supporting Information)). I_1 differs from the native conformation in that residues 76−80 are folded down to contact residues 64−67. This leaves residues 107−111 without any contacts and forces residues 68−75 into the plane immediately below the one they occupy in the native state. These residues prevent the first 50 residues, which contain substantial native structure, from interacting with the core. In I_2, the core is completely formed, but residues 74−76 displace 48−50. Residues 42−50 form an additional layer on top of the misfolded back surface (in the orientation shown in Figure 1).

At $T = 0.8$, rearrangement from I_1 and I_2 to the native is not observed within 200 × 10^6 MC steps. Consequently, we studied the transition at a higher temperature, $T = 1.0$. This temperature is slightly below $T_m \approx 1.05$, so that the native state basin is still the global free energy minimum, but the chain undergoes large fluctuations in energy (about 100 energy units) within this minimum under equilibrium conditions. For each of I_1 and I_2, we performed 20 trials that began with the intermediate and were allowed to proceed for 200 × 10^6 MC steps or until the native state was reached (Figure 4b). Of the trials that began in I_1, all 20 unfolded to at least $Q_0 = 89$ (Figure 1d), at which point the chain typically had (Q_c, Q_s) ≈ (31, 0). A majority of these went as low as $Q_0 = 49$, at which point (Q_c, Q_s) ≈ (24, 0), but the prevalence of such low Q_0 could be due to the elevated temperature. Of the trials that began in I_2, all of them unfolded completely ($Q_0 \approx 0$). Refolding proceeded directly to the native state along the fast pathway in all 20 of the trials that began in I_1 and 8 of the trials that began in I_2. The remaining 12 trials that began in I_2 failed to make any significant progress toward folding (as will be seen below, there is a free energy barrier to core formation at $T = 1.0$, so partial formation of the core in these trials is unfavorable). The 40 trials indicate that the interconversion of these intermediates requires substantial unfolding and follows essentially the same pathways as folding.

3.3. Transition State Analysis. Having mapped the reactant, intermediates, and product to the Q_cQ_s-plane, we now consider the transition states. As described above, once the complete core is formed, folding proceeds directly to the native state (or to an intermediate that is only relatively short lived). This suggests that the rate-determining transition state for correct folding precedes core formation. The multiple-path folding mechanism detailed above has several transition states associated with it (each transition state is actually an ensemble of states). In addition to the one along the "fast track" that connects the fully denatured state to the native state, we expect to find a transition state for each path connecting the fully denatured state to a long-lived intermediate and one for each path connecting an intermediate to the native state. In this section, we present a detailed description of the analysis that we use to identify structures which are representative of the transition state ensemble. Readers concerned only with the final results of the analysis may wish to skip to section 3.3.1 in which we present a summary and discuss the relation of the transition states to the rest of the kinetic analysis.

We begin our analysis of the transition states by use of the last structures at each Q_0 that make up the seven ("fast track") trajectories of the N_1 trials. We searched backwards along each such trajectory to identify the structures that have transmission coefficients[8,25] for reaching the basin around the native state (p_{N_1}) in the range $0.35 \le p_{N_1} \le 0.65$. A series of activated dynamics trials[8] beginning with each of the tested structures from the N_1 trajectories were performed; they proceeded for 5 × 10^6 MC steps or until the native state was reached. The lowest energy structure that occurred in each trial was saved and used to determine to which basin the chain had gone in that trial; in contrast to continuum models, each structure has a well-defined energy that acts as its "signature", so that it is more efficient to use the energy than to make a structural analysis. The basin around the native state was taken to include all states with energies $E < -202.0$; the native state energy is $E = -224.157$. An energy of −202.0 is lower than that of any of the long-lived intermediates and higher than that of the short-lived intermediates described above in conjunction with the "fast track" (so that minor local minima that are along the "fast track" and are close to the native state are included in the broadened native basin). On the basis of trials described below, the I_1 basin was taken to include I_1 itself with energy $E = -199.842$ and three additional states with energies $E = -199.221$; $E = -198.160$, $E = -192.906$; the I_2 basin was taken to include I_2 itself with energy $E = -200.382$ and two additional states with energies $E = -199.098$ and $E = -198.534$. For example, if the lowest energy state in a particular activated dynamics trial was $E = -212.800$, which is below $E = -202.0$, that trial was considered to have reached the basin around the native state, whereas, if the lowest energy state was $E = -198.160$, that trial was considered to have fallen into the basin around I_1. That this categorization is appropriate was confirmed by examining the contributing structures. Initially, each trajectory was searched at intervals of 20 in Q_0 (i.e., $Q_0 = 156, 136, 116, ...$) with 10 trials for each starting structure. Once the interval in which $0.35 \le p_{N_1} \le 0.65$ was bracketed (p_{N_1} increases almost monotonically with Q_0 for this set of structures), that interval was searched in intervals of 5 in Q_0 with 20 trials at each starting structure. If a structure with $0.35 \le p_{N_1} \le 0.65$ was found at this point, it was taken to be a member of the transition state ensemble for the native state. Otherwise, further searching was carried out in intervals of 1 in Q_0, again with 20 trials at each starting structure. A set of 10 or 20 trials was terminated if, regardless of the results of the remaining trials, that set could not yield $0.35 \le p_{N_1} \le 0.65$, so the total number of trials performed (1208) was not an integral multiple of 10. Moreover, in some of the original seven trajectories, more than one structure was found which had $0.35 \le p_{N_1} \le 0.65$. Because structures from the same trajectory are clearly correlated, each structure was weighted during the calculation of the average contact map of the transition state such that each original trajectory had equal weight (e.g., if there were three structures that satisfied the condition on p_{N_1} in a given trajectory, they each had a weight of one third).

In total, 15 structures were found with $0.35 \le p_{N_1} \le 0.65$. Although 15 is a negligible fraction of the actual number of configurations making up the transition state ensemble (see the Discussion for an estimate), the important structural characteristics of the ensemble are expected to be evident from this sample. This is an essential requirement for the existence of a reduced set of progress variables, such as Q_c and Q_s. As an example, a structure with $p_{N_1} = 0.45$ is shown in Figure 1e. In searching backwards along the N_1 trajectories, p_{N_1} first dropped below unity around $Q_0 = 76$ in most cases. The transition state structures range in Q_0 from 31 to 61 and in energy from -126.882 to -85.717. The average contact map of the transition state structures is shown in Figure 5Sa (Supporting Information); comparing it with Figure 3S, we see that the nontransient ($\bar{p}_{ij} \ge 0.5$) contacts are essentially a subset of the core contacts ($\bar{Q}_c = 19.1$ and $\bar{Q}_s = 1.1$). Of the $15 \times 20 = 300$ activated dynamics trials starting in transition state structures, 141 reached the basin around the native state ($E < -202.0$; of these, 87 reached the native state itself with $E = -224.157$). Thus, the average transmission coefficient for this set of structures is 0.47, which is close to the ideal value of 0.5.

It is important to determine what happens to the trajectories that do not reach the native state basin during the simulation time. Eight of the 300 trials went to the basin around I_1 (five of which went to I_1 itself), and none went to the basin around I_2 (in all 1208 trials, 40 went to the I_1 basin, of which 31 went to I_1 itself, and none went to the I_2 basin). Almost all of the rest of the 300 trials resulted in states with $-202.0 < E < -160.0$ that were each reached at most a few times. Only 25 of the 300 trials failed to reach a state with energy $E < -160.0$, which represents a reasonable upper bound for the long-lived intermediates based on the observation that much longer trials (the 50 trials of up to 200×10^6 MC steps described above or 400 additional trials that stopped at 50×10^6 MC steps but were otherwise the same in procedure as the original 50 trials) always reach $E \le -165.922$. The preponderance of lowest energy states with $E < -160.0$ in the 300 activated dynamics trials that began in transition state structures with energies in the range $-126.882 \le E \le -85.717$ indicates that these structures are committed to a rapid accumulation of native-like structure that leads to a low-energy state. The fact that a single such transition state structure can fold rapidly to either the native state or a long-lived intermediate, such as I_1 or I_2, indicates that it represents a "branchpoint".

We carried out a similar analysis for the trajectories leading to I_1 and I_2 (Figures 3c,d and 4a). The procedure differed slightly in that we restricted the initial search in intervals of 20 in Q_0 to $Q_0 \le 71$ (based on the Q_0 of the N_1 transition states: $31 \le Q_0 \le 61$) and expanded it only when necessary (i.e., if, in a given trial, the structure with $Q_0 = 71$ had $p < 0.35$, we tried structures with higher Q_0). We first describe the results of the trials that began in structures taken from the I_1 trajectories ($31 \le Q_0 \le 71$). In total, we performed 765 trials; 135 reached $E < -202.0$ (of which 92 reached the native state itself), 214 reached I_1 ($E = -199.842$, $Q_0 = 147$, $Q_c = 33$, and $Q_s = 0$), 22 reached a state with $E = -199.221$ ($Q_{I_1} = 151$ out of 163, $Q_c = 33$, and $Q_s = 1$), 32 reached a state with $E = -198.160$ ($Q_{I_1} = 144$ out of 165, $Q_c = 33$, and $Q_s = 1$), 20 reached a state with $E = -192.906$ ($Q_{I_1} = 136$ out of 162, $Q_c = 29$, and $Q_s = 1$), and none reached any state in the I_2 basin. All the other lowest energy states appeared in five or fewer trials. On the basis of their prevalence in this set of trials, we decided to consider the states with $E = -199.221$, $E = -198.160$, and $E = -192.906$ to be part of the basin around I_1 (they appear as lowest energy states in only nine of the 1208 N_1 trials and none of the I_2 trials). In analogy to the first set of activated dynamics trials described above, we calculated the probability of reaching the basin around I_1 (p_{I_1}) and took any structure with $0.35 < p_{I_1} < 0.65$ to be part of the transition state for I_1. There are 14 such structures. They range in Q_0 from 41 to 71 and in energy from -133.123 to -98.429; an example with $p_{I_1} = 0.6$ is shown in Figure 1f. Their average contact map is shown in Figure 5Sb (Supporting Information); $\bar{Q}_c = 21.851$ and $\bar{Q}_s = 0.315$. Of the $14 \times 20 = 280$ trials that began in transition state structures, only nine trials had lowest energy states with $E > -160.0$, so that, as for N_1, the transition state structures are committed to a rapid decrease in energy. Of the remaining 271 trials, 25 reached $E < -202.0$ (13 of which reached the native state itself), which confirms that individual structures in the transition state region can go to either the native state or I_1. Consistent with the observed exchange between the N_1 and I_1 pathways, there is substantial similarity between the average contact map of the I_1 transition state and that of the N_1 transition state (Figures 5Sa,b). Twenty-three of the 32 contacts with $p_{ij} \ge 0.5$ in the N_1 average contact map appear with $p_{ij} \ge 0.5$ in the I_1 map (there are 47 contacts with $p_{ij} \ge 0.5$ in the I_1 map); 17 of these 23 are in the core. To quantitate this similarity, we calculated the Pearson linear correlation coefficient for the p_{ij} of the two maps over the *ij*-pairs that correspond to the 205 contacts in the native structure, I_1, or I_2 ($r_{N_1I_1}$). We restricted the calculation to this subset of contacts because it contains all of the contacts with $p_{ij} > 0.55$ in the N_1, I_1, and I_2 transition states (all three states were included to allow direct comparison with $r_{N_1I_2}$ and $r_{I_1I_2}$ below), and any correlation would be obscured by inclusion of the rest of the contacts which are much larger in number and are always only transiently populated ($p_{ij} \approx 0$). The correlation is $r_{N_1I_1} = 0.718$, which is highly significant. As suggested above, this correlation derives primarily from contacts in the core. However, important differences do exist; in the I_1 map, the native hairpin that starts at (70−73) has already overextended to form contacts (67−76), (66−77), (65−78), and (64−79), and the large parallel sheet in the core is capped by the nonnative contact (53−80) (Figures 1f and 5Sb). The corresponding analysis of the trajectories leading to I_2 is described in Supporting Information.

In addition to the above transition states, we expect to find transition states that separate the fully denatured state from the low-energy ones. To find structures representative of the latter set, we extended the activated dynamics analysis to lower Q_0 along the N_1, I_1, and I_2 trajectories. The procedure was the same as that described above, except that, rather than monitoring the probability of reaching a particular low-energy minimum (p_{N_1}, p_{I_1}, or p_{I_2}), we monitored the probability of reaching any low-energy state with $E < -160.0$ (p_L). Structures were considered to be members of a transition state ensemble if $0.35 \leq p_L \leq 0.65$. From the trials that began in structures taken from the seven N_1 trajectories, nine structures were found that had $0.35 \leq p_L \leq 0.65$; they ranged in Q_0 from 18 to 41 and in energy from −109.472 to −59.156. An average contact map was calculated as before (such that each of the original trials had equal weight); it was found that $\bar{Q}_c = 11.571$, and $\bar{Q}_s = 0.857$. Of the $9 \times 20 = 180$ trials starting in these structures, 93 reached $E < -160.0$ (yielding $\bar{p}_L = 0.52$). For the 87 trials which failed to reach $E < -160.0$, the average final energy was $\bar{E} = -111.956$ and the average final overlap with the native state was $\bar{Q}_0 = 45.1$. In other words, because the transition states correspond to the E and Q_0 of a fully denatured disordered globule, which is the equilibrium denatured state at the given temperature ($T = 0.8$), the non-reactive trajectories do not move further towards the unfolded state. This contrasts with activated dynamics studies of simple (conventional) reactions where the transition state can be clearly distinguished from the reactants. From the trials that began in structures taken from I_1, 14 structures were found that had $0.35 \leq p_L \leq 0.65$; $16 \leq Q_0 \leq 36$, $-103.912 \leq E \leq -64.946$, $\bar{Q}_c = 8.278$, and $\bar{Q}_s = 0.722$. From the trials that began in structures taken from I_2, 10 structures were found that had $0.35 \leq p_L \leq 0.65$; $21 \leq Q_0 \leq 31$, $-112.621 \leq E \leq -79.168$, $\bar{Q}_c = 12.333$, and $\bar{Q}_s = 1.778$. In the Q_cQ_s-plane, these transition states map to the area around the leftmost branchpoint in Figure 4a. The fact that $\bar{Q}_c$ is high relative to $\bar{Q}_s$ suggests that structurally all three states consist of a further reduced version of the core. However, in general, there is not much consistent ($p_{ij} \geq 0.5$) long-range ($|i - j| > 11$) structure in these states. The few long-range contacts with $p_{ij} \geq 0.5$ are (56−83), (57−86), (97−114) in the N_1 transition state, (53−80) in the I_1 transition state, and (57−86) and (62−91) in the I_2 transition state. As before, we quantitated the pairwise similarity of the transition states by calculating the Pearson linear correlation coefficients for the p_{ij} of the 205 contacts that are in the native structure, I_1, or I_2; $r_{N_1I_1} = 0.515$, $r_{N_1I_2} = 0.639$, and $r_{I_1I_2} = 0.574$. Thus, there is clearly a significant overlap between this set of transition states, but they are not truly a common branchpoint.

3.3.1. **Summary of Transition State Analysis.** The folding trajectories and activated dynamics results shown in Figure 4a can be summarized in the simplified scheme shown in Figure 5. There are two sets of transition states, which overlap somewhat but have been separated for the purpose of description. From a random starting conformation (S), the system rapidly reaches a collapsed state with a random set of native contacts (denoted by G for "globule"), which is the equilibrium unfolded state for $T = 0.8$. The chain searches through the collapsed states for particular (primarily core) contacts to form one of the first set of transition states (described second above). Activated dynamics trials starting from any of these transition state structures, which are in the range $16 \leq Q_0 \leq 41$, have a probability of approximately 0.5 of progressing forward and eventually reaching a state with $E < -160$, or progressing backward to a state that has comparable energy and Q_0 but lacks the crucial (primarily core) contacts that are present in the transition states; i.e., the trajectory returns to the disorganized collapsed globule (G) that corresponds to the equilibrium unfolded state. On average, these states map to the earliest branchpoints in Figure 4a at $\bar{Q}_c \approx 10$ and $\bar{Q}_s \approx 1$.

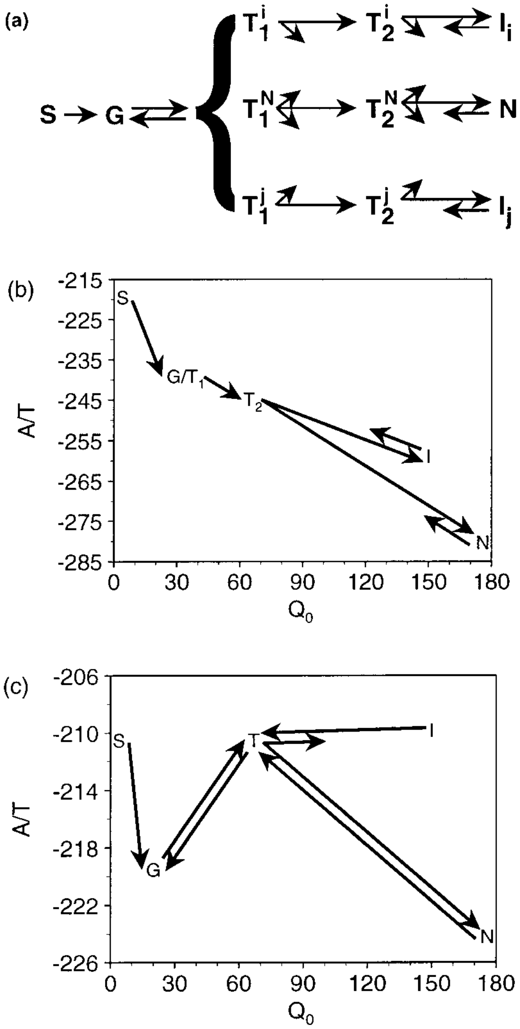

Figure 5. Schematic description of the 125-mer folding reaction. (a) Representative pathways, where S is a random starting structure, G is a disordered globule, T are transition states, N is the native state, I_i is intermediate i, and $I_{j\neq i}$ includes all other long-lived intermediates; diagonal arrows indicate that a state can lead to either of the other pathways (for example, T_1^i leads predominantly to T_2^i but can also lead to T_2^N and $T_2^{j\neq i}$). (b) Free energy at $T = 0.8$. (c) Free energy at $T = 1.0$.

If the chain progresses forward (Q_0 increases), it almost immediately (prior to actually reaching $E < -160$) encounters the second set of transition states ($31 \leq Q_0 \leq 86$), at which point it becomes committed to a rapid accumulation of primarily native structure to yield a low-energy state ($E < -160$). However, the chain is only committed to a particular pathway with a probability of 0.5. In other words, a member of the T_2^N ensemble has a probability of 0.5 of reaching the native basin and a probability of 0.5 of becoming trapped in one or another of the basins around other low-energy states, such as I_1, I_2, or I_{3+} (denoted I_i and $I_{j\neq i}$ in Figure 5). Similarly, a member of the T_2^i ($i = 1$ or 2 in the previous section) ensemble has a probability of 0.5 of reaching the basin around I_i and a probability of 0.5 of either becoming trapped in one or another of the basins around other low-energy states ($I_{j\neq i}$) or reaching the basin around the native state. It should be noted that, in reality, the situation is slightly more complex in that each transition state exhibits preferences for certain pathways; for example, in the analysis detailed above, it was found that T_2^N tends to misfold to I_1 but not to I_2. This is in accord with the fact that there is more structural similarity between T_2^N and T_2^1 than between T_2^N and T_2^2. Once the chain has moved off the original pathway (for example, from N_1 to I_1 via T_2^N), it must break stabilizing (native and nonnative) contacts to rearrange to the original target state, which is slow at $T = 0.8$. In the Q_cQ_s-plane, the second set of transition states coincides with the branchpoints at $Q_c \approx 20$ in Figure 4a. From the complexity of Figure 3, relative to Figure 4a, it is evident that the scheme in Figure 5 is oversimplified. Additional protein coordinates, other than Q_c and Q_s, would be required for a more detailed quantitative description of the kinetics and the multiplicity of pathways. Nevertheless, the essential features of 125-mer folding are captured in this reduced representation and the scheme presented in Figure 5.

The above exploration of the transition states indicates that the search for the native state can be reduced to the search for structures that contain the core contacts in the absence of other contacts that slow folding. Earlier analysis indicated that the core is formed by a random search that is restricted by low energy, kinetically accessible, cooperative native contacts.[22] Confirmation of this idea is given in Supporting Information.

4. Thermodynamic Results

As described in Supporting Information, we obtained the density of states of the accessible conformational space as a function of E, Q_c, and Q_s [$\omega(E, Q_c, Q_s)$] from Monte Carlo sampling techniques. A sum over Q_c and Q_s to yield $\omega(E)$ is presented in Figure 7S (Supporting Information) for comparison with densities of states of other models. In contrast to the corresponding curves for 27-mers with random interactions,[35] $\omega(E)$ of the 125-mer is clearly non-Gaussian. This shift towards an essentially exponential dependence ($\log_{10} \omega$ increases almost linearly with E) derives from the optimization of the sequence,[24] and, as will be seen, has consequences for the thermodynamics.

From the density of states, we can calculate the free energy, energy, and entropy as functions of the progress variables at any temperature. Results for the temperature of the majority of the kinetic trials ($T = 0.8$) and for the temperature of the trials that monitored the interconversion from the intermediates I_1 and I_2 to the native state ($T = 1.0$) are presented; results for temperatures above T_m ($T_m \approx 1.05$) are presented elsewhere.[36] We plot A/T, E/T, and S rather than A, E, and TS, to allow direct comparison between the two different temperatures. Because the overall changes in the thermodynamic quantities are large in comparison to many of the features of interest (barriers and minima), the latter can be difficult to identify. To aid in the analysis, we show three sets of plots. Three-dimensional surfaces are shown in Figure 6, and contour plots are shown in Figure 8S (Supporting Information); the three-dimensional and the contour plots are colored identically and go from blue (low values) to red (high values). The third set of plots (Figure 7) shows the free energy, energy, and entropy along the N_1 and I_1 average trajectories shown in Figure 4a parameterized in terms of Q_0; in other words, we calculate $\bar{Q}_c(Q_0)$ and $\bar{Q}_s(Q_0)$ from the last structures at each Q_0 and plot $A[Q_0, \bar{Q}_c(Q_0), \bar{Q}_s(Q_0)]/T$, $E[Q_0, \bar{Q}_c(Q_0), \bar{Q}_s(Q_0)]/T$, and $-S[Q_0, \bar{Q}_c(Q_0), \bar{Q}_s(Q_0)]$. Corresponding curves for the I_2 average trajectory are not shown because, as discussed below, I_2 is not sufficiently well separated from the N_1 pathway in the Q_cQ_s-plane (Figure 4a).

The structures obtained during the Monte Carlo sampling allow decomposition of the density of states into terms of variables other than Q_c and Q_s for the calculation of equilibrium averages. In Figures 8 and 9S (Supporting Information), we show the average number of contacts ($\langle C\rangle$), the average number of native contacts ($\langle Q_0\rangle$), the average number of contacts in antiparallel sheets ($\langle C_{as}\rangle$), and the average number of contacts in parallel sheets ($\langle C_{ps}\rangle$) at $T = 0.8$ and $T = 1.0$, respectively. The average number of contacts in helices is not shown because it is nonnegligible only in the fully denatured region and never exceeds 10. The values of C, Q_0, C_{as}, and C_{ps} for the structures discussed in the kinetic analysis are given in the caption to Figure 1. In what follows, we describe the relations between the thermodynamic quantities (Figures 6 and 7), the equilibrium averages (Figures 8 and 9S), and the kinetic behavior.

At the lower temperature ($T = 0.8$), the free energy slopes smoothly downward overall towards the native state at (34, 15); $A(0, 0)/T - A(34, 15)/T = 46.5$ (Figures 6a and 8Sa). Even though (0,0) is the highest point on the energy surface (Figure 6b), it is not the highest point on the free energy surface (Figure 6a) because the large entropy at (0,0) counterbalances the energy. The random starting configurations of the kinetic trials, as described above, are very open structures which fall in the range $0 \leq Q_0 \leq 12$, $0 \leq Q_c \leq 4$, and $Q_s = 0$. These are highly excited states with energies ($-31 \leq E/T \leq -6$) which are far above the average in the corresponding region of the Q_cQ_s-plane ($E \approx -112$). Thus, the initial nonspecific collapse ($<10^6$ MC steps) takes the system from a point above the equilibrium energy surface shown in Figure 6b onto it [dividing the energies observed for the disordered globule in the kinetic trials (see section 3.1) by $T = 0.8$, we have $-169 \leq E/T \leq -100$, $0 \leq Q_c \leq 16$, and $0 \leq Q_s \leq 3$]. In this collapse, the number of contacts increases from $15 \leq C \leq 75$ to close to the equilibrium average, $\langle \mathrm{C}\rangle \approx 115$ (Figure 8b), but the chain remains disordered with $\langle Q_0\rangle \approx 30$ (Figure 8a). There is a relatively small amount of secondary structure: $\langle C_{as}\rangle \approx 35$ and $\langle C_{ps}\rangle \approx 25$ (compare Figure 8c and d to the values given in the caption to Figure 1). The free energy is nearly flat in this region, which covers the area within the triangle defined by coordinate pairs (Q_c, Q_s) = (3, 0), (16, 0), and (7, 5). (Figure 8Sa). This virtual plateau in the free energy is a consequence of the compensation between the energy and the entropy. Although the energetic and entropic components exhibit plateaus in the region defined by (Q_c, Q_s) = (0, 0), (10, 0), and (0, 8), they change rapidly in a competing manner in the flat region of A/T [compare the behavior around (10,1) in Figure 6a to that in Figure 6b and c and that in Figure 8Sa to that in Figure 8Sb and Sc]. The contrast in the behavior of the free energy and those of its components is particularly clear in the plots in which the thermodynamic quantities are shown as functions of Q_0 (Figure 7a−c); for example, at $Q_0 \approx 30$, A/T is flat while both E/T and $-S$ change steadily. This is a striking example of the compensation between energy and entropy, which is an important aspect of the folding reaction. As the number of native contacts increases, the energetic component of the free energy drops due to the stabilizing interactions, but the entropic component rises equally rapidly due to the concomitant decrease in the number of accessible states.

At higher Q_c and Q_s, we find two major minima on the free energy surface (Figures 6a and 8Sa). The first of these is the native state global minimum at (34,15) which corresponds to the product of the folding reaction. The second of these is a

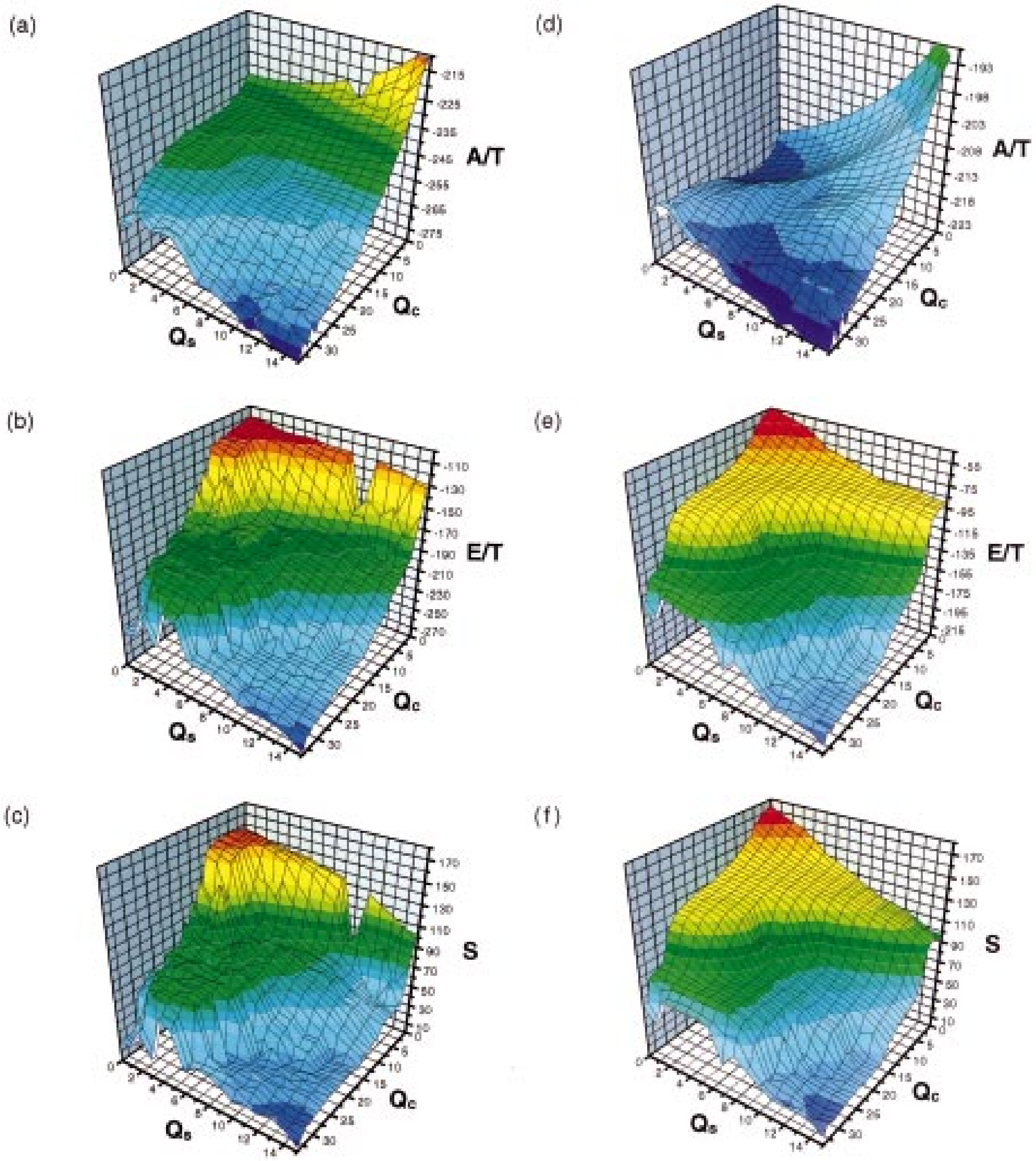

Figure 6. Surface plots of the reaction thermodynamics. The normalized free energy (A/T) at (a) $T = 0.8$ and (d) $T = 1.0$ with the respective (b and e) energetic (E/T) and (c and f) entropic (S) components. The gridlines on the surfaces are spaced in intervals of $\Delta Q_c = 1$ and $\Delta Q = 1$. The thermodynamic values at (34,0), (34,1), (34,2), and (34,3) were set to those at (33,0), (33,1), (33,2), and (33,3), respectively, to make the minimum around I_1 more easily visible. The plots were made with the program Origin.

local minimum at (33,0) and corresponds to the intermediate I_1. Both of these minima are stabilized by the energetic component of the free energy and destabilized by its entropic component; they are minima in Figures 6b and 8Sb and maxima in Figures 6c and 8Sc. We focus first on the global minimum at (34,15). Although this coordinate pair contains many states other than the native one, the average values of the structure variables are close to those for the native conformation: $\langle C\rangle \approx 172$, $\langle Q_0\rangle \approx 172$, $\langle C_{as}\rangle \approx 80$, and $\langle C_{ps}\rangle \approx 74$ (Figure 8); for the native structure, $C = 176$, $Q_0 = 176$, $C_{as} = 81$, and $C_{ps} = 76$ (Figure 1a). The minimum at (33,0) is markedly higher in free energy than the global minimum [$A(33, 0)/T - A(34, 15)/T = 18.5$] due primarily to the energetic component [$E(33, 0)/T - E(34, 15)/T = 32.2$, while $S(33, 0) - S(34, 15) = 13.7$]; the difference in energy corresponds closely to that between I_1 and the native state [$\Delta E/T = (-199.842 + 224.157)/0.8 = 30.4$]. At (33,0), the average values of the structure variables are close to the corresponding I_1 values: $\langle C\rangle \approx 158$, $\langle Q_0\rangle \approx 141$, $\langle C_{as}\rangle \approx 82$, and $\langle C_{ps}\rangle \approx 55$ (Figure 8), while, for I_1, $C = 162$, $Q_0 = 147$, $C_{as} = 84$, and $C_{ps} = 57$ (Figure 1b). Of particular interest is the low value of $\langle C_{ps}\rangle$ relative to that of the native state ($C_{ps} = 76$). Overall, $\langle C_{ps}\rangle$ is much flatter than the other structural averages at $T = 0.8$ (compare Figure 8d to Figure 8a−c). Formation during folding of antiparallel sheet contacts at the expense of parallel sheet contacts leads to intermediates with high C_{as} and low C_{ps}. A preference for antiparallel sheet contacts

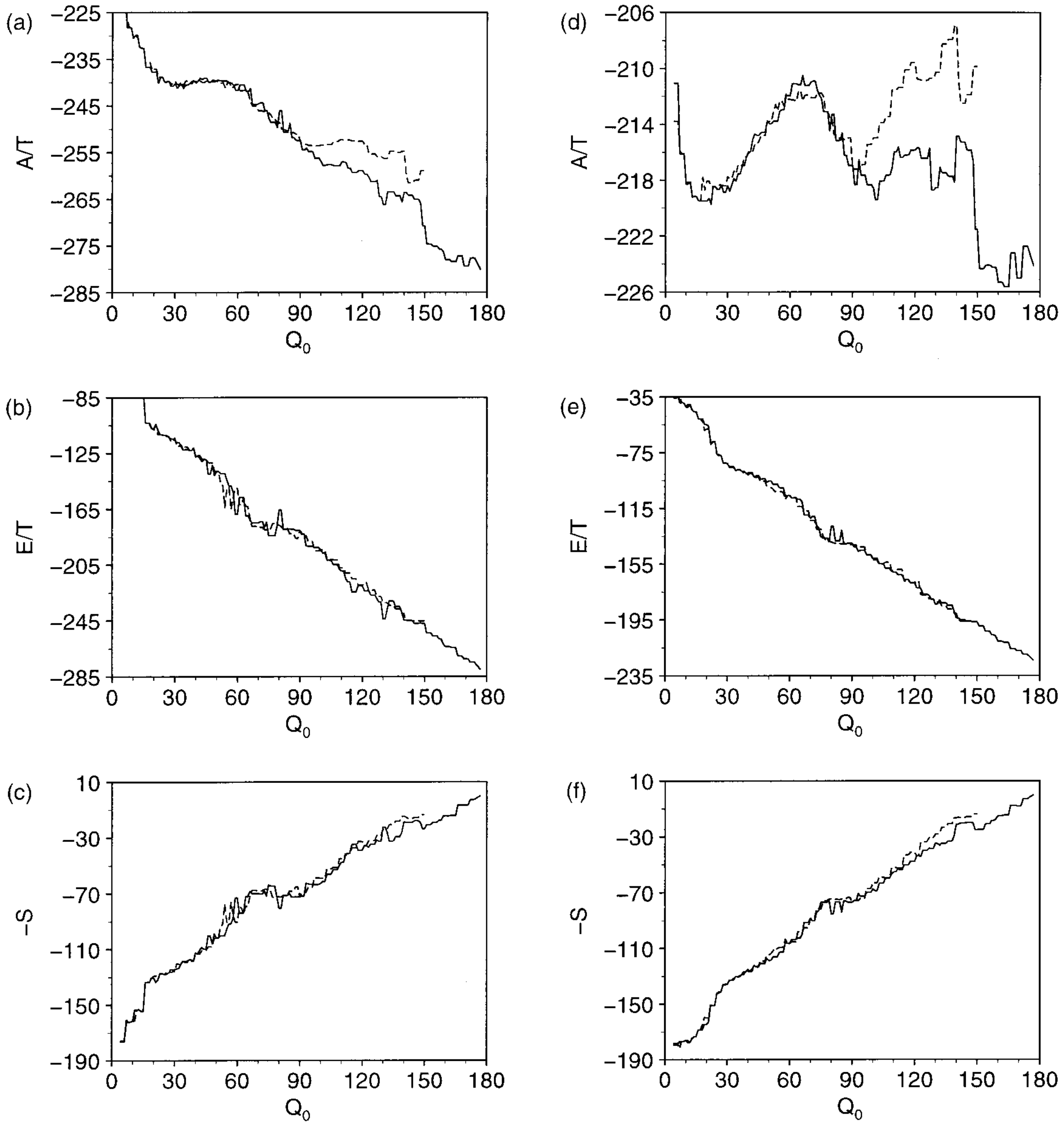

Figure 7. Reaction thermodynamics decomposed into functions of Q_0, Q_c, and Q_s and plotted for the coordinate triplets along the average trajectories of the last structures at each Q_0 [Q_0, $\bar{Q}_c(Q_0)$, and $\bar{Q}_s(Q_0)$]; (solid lines) N_1 trajectory and (dashed lines) I_1 trajectory. The free energy (A/T) at (a) $T = 0.8$ and (d) $T = 1.0$ with the respective (b and e) energetic (E/T) and (c and f) entropic ($-S$) components.

over parallel ones is consistent with the results of the statistical analyses of 125-mer folding,[22,24] in which it was found that the number of antiparallel sheet contacts in the native structure correlated strongly with folding ability, but that the number of parallel sheet contacts did not. In the earlier studies,[22,24] this correlation was explained in terms of the fact that antiparallel sheets can be initiated more readily by a random search than can parallel sheets because the former are typically shorter ranged (to form a parallel sheet, the chain must loop around). The I_1 minimum is separated from the native one by more open structures with fewer antiparallel sheet contacts; for example, at (33,4), $\langle C\rangle \approx 138$, $\langle Q_0\rangle \approx 106$, $\langle C_{as}\rangle \approx 59$, and $\langle C_{ps}\rangle \approx 52$ (Figure 8). The idea that interconversion from I_1 to the native state requires breaking antiparallel sheet contacts that preempt parallel sheet contacts is consistent with both the structure of the I_1 transition state (Figures 1f and 5Sb), in which the antiparallel sheet starting at (70−73) has already overextended, and with the structures observed during interconversion at $T = 1.0$ (Figure 1d).

Although there is a small local minimum in the free energy at the coordinates of the intermediate I_2 ($Q_c = 34$ and $Q_s = 11$), it does not derive from I_2. At (34,11), there are states with energies as low as $E = -217$ ($E/T = -271.25$), and these, rather than I_2, dominate the Boltzmann-weighted averages. The average energy ($\langle E\rangle/T \approx -263.251$) is much lower than the energy of I_2 ($E/T = -250.478$), and $\langle Q_0\rangle$ ($\langle Q_0\rangle \approx 161$) is much higher than that of I_2 ($Q_0 = 150$). The other structural averages ($\langle C\rangle \approx 165$, $\langle C_{as}\rangle \approx 79$, and $\langle C_{ps}\rangle \approx 67$) are close to the values for I_2 ($C = 164$, $C_{as} = 76$, and $C_{ps} = 68$), but these measures are less specific, so the correspondence (without concomitant correspondences in E and Q_0) may simply be a reflection of general features of the native-like states. As discussed in section 3.3, most of the states with $E < -202.0$ interconvert readily to the native state, so that the minimum at (34,11) can be viewed as a local feature in the broader native state basin. This is not to say that I_2 is not an energetically stabilized local free energy minimum, simply that its visualization as such would require separation into a third dimension with a coordinate that better separated N_1 and N_2 from I_2.

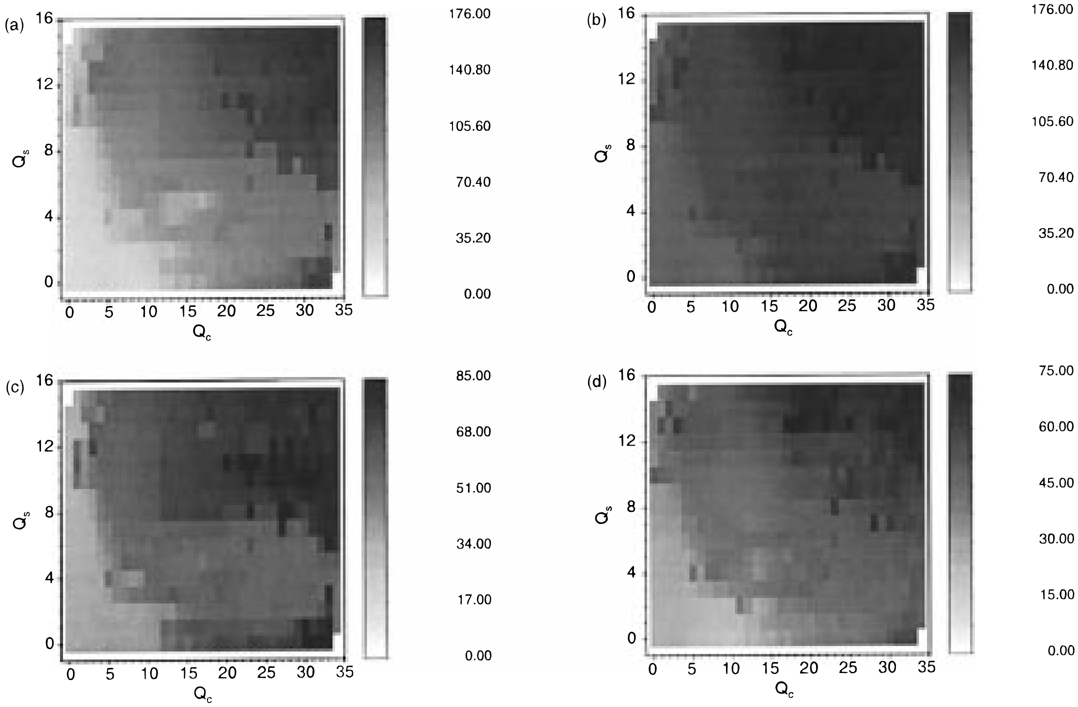

Figure 8. Density plots of equilibrium averages at $T = 0.8$. (a) Average contact overlap with the native state ($\langle Q_0 \rangle$), (b) average number of contacts ($\langle C \rangle$), (c) average number of contacts in antiparallel sheets ($\langle C_{as} \rangle$), and (d) average number of contacts in parallel sheets ($\langle C_{ps} \rangle$). In each case, the maximum value on the scale is the highest observed rounded up to the nearest integer.

At the higher temperature ($T = 1.0$), the entropic component, as expected, plays an even more significant role in the free energy than at $T = 0.8$. As already mentioned, $T = 1.0$ is below $T_m \approx 1.05$, so that the native basin is still the global free energy minimum. In the fully denatured region ($0 \le Q_c \le 16, 0 \le Q_s \le 8$), the entropic component (Figures 6f and 7f) decreases more than the energy increases (Figures 6e and 7e). This creates a broad minimum around (6,0) (Figures 6d and 8Sd): $A(6, 0)/T - A(34, 15)/T = 4.4$, $E(6, 0)/T - E(34, 15)/T = 148.0$, and $-S(6, 0) + S(34, 15) = -143.6$. Consistent with the entropic nature of this minimum, the chain is relatively open ($\langle C \rangle \approx 87$) and unstructured ($\langle Q_0 \rangle \approx 22$, $\langle C_{as} \rangle \approx 28$, and $\langle C_{ps} \rangle \approx 17$) (Figure 9S). At higher Q_c, most of the states are well structured (low entropy), and the free energy surfaces at the two temperatures are more similar. The minimum at (33,0) is still a local minimum (Figures 6d and 8Sd), but there is a much larger contribution to the averages from structures other than I_1; at (33,0), $\langle C \rangle \approx 136$, $\langle Q_0 \rangle \approx 107$, $\langle C_{as} \rangle \approx 66$, and $\langle C_{ps} \rangle \approx 58$ (Figure 9S and caption to Figure 1b). Although the I_1 minimum is stabilized somewhat more relative to the native state in comparison with $T = 0.8$ [$A(33, 0)/T - A(34, 15)/T = 11.0$, $E(33, 0)/T - E(34, 15)/T = 55.7$, $-S(33, 0) + S(34, 15) = -44.7$ versus $A(33, 0)/T - A(34, 15)/T = 18.5$, $E(33, 0)/T - E(34, 15)/T = 32.2$, and $-S(33, 0) + S(33, 0) = -13.7$], it is destabilized relative to the fully denatured state (Figure 7d). This accounts for the observation that the chain unfolds to interconvert to the native state (within 200×10^6 MC steps) at $T = 1.0$ but not at $T = 0.8$. At higher Q_s, the native state at (34,15) is destabilized relative to other states in the "native basin" ($E < -202.0$), and the global free energy minimum shifts to (34,13) (Figure 6d); $A(34, 13) - A(34, 15) = -1.5T$, $E(34, 13) - E(34, 15) = 8.2T$, $-TS(34, 13) - TS(34, 15) = -9.7T$. This destabilization was not a significant factor in the kinetic trials described above that monitored interconversion from the intermediates I_1 and I_2 to the native state (and stopped upon first passage) because the native state at (34,15) can be reached easily by local excursions from the global minimum at (34,13). The structural averages at (34,13) are $\langle C \rangle \approx 164$, $\langle Q_0 \rangle \approx 159$, $\langle C_{as} \rangle \approx 78$, and $\langle C_{ps} \rangle \approx 66$.

4.1. Transition States. We now consider the behavior of the equilibrium averages in the regions that contain the transition states. We begin with the higher temperature ($T = 1.0$) results because they are more straightforward. As described in section 3.2, at this temperature, the chain either folds completely or fails to make any appreciable progress towards folding. Thus, it is likely that, at $T = 1.0$, there is only a single set of transition states, which falls at roughly the Q_0, Q_c, and Q_s of the second set of transition states for the $T = 0.8$ folding reaction (T_2 in Figure 5). The states contributing to T_2^N fall at ($31 \le Q_0 \le 61$, $\bar{Q}_c = 19.1$, $\bar{Q}_s = 1.1$). At $T = 1.0$, this region is part of a broad barrier that separates the denatured minimum at (6,0) from the intermediate and native minima at (33,0) and (34,13), respectively (Figures 6d and 8Sd); the thermodynamic quantities are $A(19, 1)/T - A(6, 0)/T = 4.3$, $E(19, 1)/T - E(6, 0)/T = -21.9$, $-S(19, 1) + S(6, 0) = 26.2$. The behavior is particularly clear in Figure 7; the barrier, which falls between $Q_0 \approx 20$ and $Q_0 \approx 80$ along the N_1 pathway derives from an increase in $-S$ (Figure 7f) without a comparable decrease in E/T (Figure 7c). Consistent with the lag in the energetic component, the structural averages at (19,1) are almost those of the fully denatured state: $\langle C \rangle \approx 99$, $\langle Q_0 \rangle \approx 38$, $\langle C_{as} \rangle \approx 36$, and $\langle C_{ps} \rangle \approx 23$ (Figure 9S). At the Q_0 values that follow the transition state region ($Q_0 > 61$), the energy drops sharply (Figure 7e), and the free energy decreases (though less than the energy due to an increase in the entropic component of the free energy) (Figure 7d). The structural averages are markedly higher after the barrier (compare Figure 9S to Figure 8Se,f); for example, at (27,6), the nearest integer coordinate pair to $\bar{Q}_c(Q_0 = 80) = 27.1$ and $\bar{Q}_s(Q_0 = 80) = 5.7$, $\langle C \rangle \approx 126$, $\langle Q_0 \rangle \approx 94$, $\langle C_{as} \rangle \approx 51$, and $\langle C_{ps} \rangle \approx 46$. Thus, at the higher temperature ($T = 1.0$), the set of transition states associated with the N_1 path coincides with a barrier that derives from a lag in the energetic component of the free energy relative to the entropic component.

The corresponding set of transition states for the I_1 trajectories falls at ($41 \leq Q_0 \leq 71$, $\bar{Q}_c = 21.8$, $\bar{Q}_s = 0.3$). Since they are very close to the N_1 transition states, their thermodynamic description is essentially the same. The transition states for the I_2 trajectories fall at ($49 \leq Q_0 \leq 86$, $\bar{Q}_c = 22.0$, $\bar{Q}_s = 4.9$). In spite of the increase in Q_s, these transition states encounter the same barrier in the free energy at $T = 1.0$. However, as can be seen from the clustering of the contour lines around (22,5) in Figure 8Se,f, E/T and $-S$ change much more sharply in this region of the Q_cQ_s-plane. At (22,5), $\langle C\rangle \approx 117$, $\langle Q_0\rangle \approx 73$, $\langle C_{as}\rangle \approx 43$, and $\langle C_{ps}\rangle \approx 37$. As expected, from the higher range of Q_0 identified kinetically, the I_2 transition states fall in a region in which the equilibrium averages are closer to their native values. This shift towards the native state is consistent with the lower prevalence of this intermediate; i.e., since the transition state is more native-like, it is harder to find, which slows its formation and decreases the likelihood of forming the associated product I_2.

We now turn to the lower temperature ($T = 0.8$) results, where there are two sets of transition states that overlap somewhat but have been separated for the purpose of description. We continue with the T_2 transition states, the coordinates of which are given immediately above. At the lower temperature ($T = 0.8$), no significant barrier is observed in these regions. Rather, the transition states ($31 \leq Q_0 \leq 71$) immediately precede an increase in the magnitude of the slope of A/T (Figure 7a). Like the free energy, the structural averages fall on plateaus of intermediate values (Figure 8): at (19,1), $\langle C\rangle \approx 136$, $\langle Q_0\rangle \approx 78$, $\langle C_{as}\rangle \approx 61$, and $\langle C_{ps}\rangle \approx 36$. The profiles along the I_2 trajectory (not shown) are almost identical to those along the N_1 trajectory but, as discussed above, this similarity could be due in part to incomplete separation of the two sets of trajectories in the Q_cQ_s-plane. Thus, the later transition states, which were identified by their behavior in activated dynamics simulations, are not transition states in the conventional sense at $T = 0.8$; i.e., they are not maxima in the free energy. Instead, they are essentially plateaus in the free energy that correspond to balanced changes in the relative importance of the energetic and entropic components. The plateau corresponds to the low-temperature limit of a free energy barrier that results from an interplay of energy and entropy; i.e., at only slightly higher temperatures, the entropy of the denatured state is sufficiently important to make the plateau into a barrier. Although it has been suggested that the absence of a free energy barrier can be due to the "wrong" choice of coordinates,[12,27] it appears to us that this is not true in the present case.

As shown in the kinetic analysis (section 3), there exists an earlier set of transition states that separates the fully denatured states from low-energy states (T_1 in Figure 5). For the N_1 transition states, $18 \leq Q_0 \leq 41$, $\bar{Q}_c = 11.6$, and $\bar{Q}_s = 0.8$; for the I_1 transition states, $16 \leq Q_0 \leq 36$, $\bar{Q}_c = 8.3$, and $\bar{Q}_s = 0.7$; for the I_2 transition states, $21 \leq Q_0 \leq 31$, $\bar{Q}_c = 12.3$, and $\bar{Q}_s = 1.8$. At the lower temperature ($T = 0.8$), as was the case with the T_2 transition states, these correspond to a plateau in the free energy (Figure 7b and c). Their structural averages are above those of the fully denatured states; for example, at (12,1), $\langle C\rangle \approx 129$, $\langle Q_0\rangle \approx 64$, $\langle C_{as}\rangle \approx 58$, and $\langle C_{ps}\rangle \approx 30$. As described above in conjunction with I_1, the accumulation of parallel sheet contacts lags behind that of antiparallel ones (Figure 8c and d).

4.2. Summary of Thermodynamic Results. At the two temperatures considered ($T = 0.8$ and $T = 1.0$), the energy and entropy surfaces exhibit strikingly similar behavior; they are high in the region of the disordered fully unfolded state and low in the regions of the native state and of the long-lived intermediates. At the lower temperature ($T = 0.8$), the energy dominates, and the free energy suface leads smoothly downhill to either the native state or a low-energy intermediate in which antiparallel sheet contacts have preempted the formation of native parallel ones. The intermediates correspond to local minima on the free energy surface that are stabilized by the energetic component. The transition regions coincide with relatively flat portions of the free energy surface that involve balanced changes in the energetic and entropic components. At the higher temperature ($T = 1.0$), the free energy surface as a whole is relatively flat and the chain folds from an entropic (denatured) minimum to an energetic (native) one. There is no trapping in specific low-energy intermediates because they are destabilized relative to the unfolded state. The transition regions coincide with a broad barrier that derives from a lag in the change in energy relative to the change in entropy. Overall, the results at both temperatures present striking examples of the compensation between energy and entropy, which is an important aspect of protein folding reactions (see also Figure 4 of ref 20). As the system collapses and folds, the entropic component of the free energy ($-TS$) rises rapidly due to the decrease in the number of states; simultaneously, the energetic component drops rapidly due to the accumulation of stabilizing contacts. Thus, the balance between these two rapidly varying quantities is the essential element in determining the free energy surface. This is very different from most small molecule reactions which are dominated by the variation of the energy along the reaction coordinate.

5. Discussion

A full description of a protein folding reaction requires an understanding of both its thermodynamics and its kinetics. The former involves a knowledge of the free energy, energy, and entropy surfaces as a function of suitable coordinates, while the latter requires a description of the motions on those surfaces that take the system from the denatured state to the native state. Because even a small protein has thousands of degrees of freedom, explicit consideration of all possible configurations is not feasible, and it is necessary to group states by projecting the configuration space onto a small number of (reaction) coordinates that characterize the system. Since the entire macromolecular system participates in the reaction and entropy plays a substantial role, it is not straightforward to find an appropriate reduced set of coordinates. These difficulties have led to the introduction of simplified models, for which one can obtain many folding (reactive) and nonfolding (nonreactive) trajectories that can be used to determine the reaction paths. In the present study, we considered a 125 residue chain subject to Monte Carlo dynamics on a simple cubic lattice. This chain has a length and a ratio of surface to buried residues in the native state that are comparable to those of "larger" well-studied proteins, such as lysozyme, barnase, and myoglobin, and the model exhibits many of the complexities observed experimentally in the folding mechanisms of these proteins. We thus expected (and found) the model to yield insights that supplement the experimental data on how such proteins fold.

A statistical analysis of two hundred 125-mer sequences demonstrated that folding is promoted by increases in the overall stability, cooperativity, and kinetic accessibility of the native state and is hindered by over-stabilization of the contacts between surface residues.[22,24] Analysis of a few sequences revealed that these correlations derived from a mechanism which involved two slow steps on the average.[22] In the first step, the chain makes a random search for a core that drives folding either directly to the native state or to a long-lived native-like intermediate. Increases in the number of stable, kinetically accessible cooperative native contacts accelerate folding by helping the chain restrict its search for the core. In the second step, the chain rearranges from the intermediate to the native state by breaking (both native and nonnative) contacts between primarily surface residues so that they can condense in correct relation to the core. Stabilization of these contacts makes them more difficult to break and thus slows folding.

To obtain a better understanding of the folding reaction of the 125-mer, we made a detailed study of a representative sequence. The sequence was chosen because both core formation and rearrangement from an intermediate are slow, but it is still capable of folding in most trials (of $\sim 10^8$ MC steps). Its native structure is composed entirely of sheets, with 76 contacts in parallel sheets and 81 contacts in antiparallel sheets (which include the turns). The folding mechanism begins with a relatively fast collapse to a disordered globule that contains a significant amount of transient secondary structure. The chain then makes a random search for the core contacts, which accumulate gradually. When about 20 (out of 34) core contacts have been formed, the chain reaches a branchpoint. In about 15% of the trials, it finds several more core contacts and folds either directly to the native state or to a short-lived intermediate (in which the last 25 residues are misfolded) that interconverts rapidly to the native state (the N_1 “fast track”). In 15% of the trials, the chain forms particular nonnative contacts by over-extending two sheets in the core, which directs the chain to a long-lived intermediate that must partially unfold to interconvert to the native state (the I_1 “slow track”). In 25% of the trials, the chain forms contacts outside the core (native and nonnative) that direct it to other long-lived intermediates (the I_2 and I_{3+} “slow tracks”). In the remaining 45% of the trials, the chain again becomes partially misfolded but succeeds in escaping from the traps within the simulation time (the N_2 trajectories). In general, the intermediates tend to lack parallel sheet contacts, which tend to be longer ranged and thus are harder to find by a random search. They differ qualitatively from “intermediates” reported for shorter lattice models (36-mers), which occur before the formation of the nucleus and are thus relatively unstructured and short lived (10^5 MC steps or about 10% of the total folding time).[37−39]

Although in the case of shorter chains, such as the 27-mer, the reaction was described adequately with a single progress variable (the number of native contacts, Q_0),[20] for the multiplicity of pathways found for the 125-mer, it was necessary to introduce a higher dimensional representation. In analogy to earlier studies of simpler reactions,[8] we used the smallest reduced set of coordinates (only two were required) that was sufficient to reflect the major structural differences between reactive (folding) and nonreactive (nonfolding) trajectories. We took the first coordinate (Q_c) to be a subset of native contacts that is present with high probability immediately prior (in Q_0) to core formation in folding trajectories and the second coordinate (Q_s) to be a subset of native contacts that appears with higher probability in folding trajectories than in nonfolding trajectories at values of Q_0 corresponding to the misfolded intermediates. Thus, in spite of the fact that, by itself, Q_0 does not provide a satisfactory description of the reaction, it was useful for determining additional progress variables once the diverse trajectories were grouped appropriately. The coordinates Q_c and Q_s were sufficient to separate (for the most part) the “fast track” (N_1) from the “slow tracks” (particularly those leading to I_1), so that the intermediate and transition state regions could be delineated and the thermodynamic behavior along the pathways could be determined.

Due to the complexity of the folding mechanism, there are several transition states (each of which is actually an ensemble of states). To identify these states, we used activated dynamics simulations on subsets of the structures sampled during folding from random configurations. For each pathway (N_1, I_1, and I_2), we identified two transition states, which overlap somewhat but have been separated for the purpose of discussion: the first separates the fully denatured (random coil) state from the low-energy (partially or fully structured) states and the second separates one low-energy basin from another. Trials starting in the first set of states either progress forward, eventually reaching a state with $E < -160$, or progress backward to a state that has comparable energy and Q_0 but lacks the crucial (primarily core) contacts that are present in the transition states. Trials starting in the second set of states always go toward a state with more overall similarity to the native state, as measured by Q_0, but about half the time, this state interconverts less readily to the product state associated with that pathway (N_1, I_1, or I_2) than does the transition state in question, so that the reaction has effectively progressed backwards. Structurally, the transition states that lead directly to the basin around the native state have roughly 20 high probability contacts, almost all of which fall within the core. The transition states that lead to the two most prevalent intermediates (I_1 and I_2) overlap significantly with that of the native state but differ in that particular native and nonnative contacts outside the core are formed with high probability. These contacts, many of which fall in nonnative antiparallel sheets that prevent formation of native parallel sheets, direct the chain to the intermediates rather than to the native state. In the Q_cQ_s-plane, the transition states correspond to branchpoints in the average pathways; the first set falls at $Q_c \approx 10$ (out of 34) and $Q_s \approx 1$ (out of 15) and the second set falls at $Q_c \approx 20$ and $0 \leq Q_s \leq 5$.

From the equilibrium sampling of the accessible configuration space and the observed kinetics, it is possible to estimate the number of conformations that compose the transition state ensembles. As discussed above, those that immediately precede the native state (T_2^N) fall at $31 \leq Q_0 \leq 61$, $\bar{Q}_c = 19.1$, and $\bar{Q}_s = 1.1$. At the average value of Q_0 ($\bar{Q}_0 \approx 45$) for the 15 representative transition state structures identified in the activated dynamics trials, the entropy is 104.950; the number of accessible states at these coordinates can be estimated as $\exp[S/k_B] \sim 10^{45}$. The mean first time of reaching $Q_c = 19$ is 1.11×10^6 MC steps, and the mean last time of reaching $Q_c = 19$ is 22.45×10^6 MC steps, so it is reasonable to take the mean time of reaching the transition state as $\sim 10^7$ MC steps. However, only a small fraction of these steps yield new structures because the acceptance rate is low (about 13% during the early part of the simulations) and many of the accepted steps will revisit structures that were already sampled. Thus, the number of structures sampled prior to reaching the transition state is $\sim 10^5$. These estimates suggest that the number of transition states is $\sim 10^{45}/10^5 = 10^{40}$. Thus, there appear to be a very large number of transition states compared, for example, to the $\sim 10^3$ estimated for a 27-mer with random interactions. The increase stems from both the increase in length of the chain (with which the number grows exponentially) and the optimization of the sequence, which shifts the transition state towards the denatured state in accord with the Hammond postulate.[40,41] This shift is an important element of the nucleation (core formation) and sequential growth mechanism that allows the chain to find the transition state in a reasonable amount of time. For off-lattice models and proteins, in which large-scale diffusive motions are possible, such a reduction in conformational space could be obtained alternatively by early formation of helices, as in the diffusion−collision model.[42−44]

At the temperature employed in the majority of kinetic trials ($T = 0.8$), the chain folds down-hill on a relatively smooth free energy surface to the low-energy minima, that consist of the native state and the intermediates. The transition regions coincide with relatively flat portions of the free energy surface that involve balanced changes in the energetic and entropic components. At slightly higher temperatures, such as the one employed to monitor interconversion from the intermediates to the native state ($T = 1.0$), the entropic component stabilizes the more open denatured state so that it becomes a free energy minimum which is separated from the native and intermediate minima by a broad barrier that coincides with a branchpoint (transition state) of the folding mechanism. This barrier derives from the fact that the energy decreases more slowly than the entropy. The free energy surface and its temperature dependence are a striking demonstration of the importance of the entropy in protein folding, which contrasts with the dominance of the energy in reactions of small molecules.

The progress variables Q_c and Q_s used in the present study are structural order parameters that were chosen based on an analysis of the mechanism specific to the 125-mer folding reaction. As a result, they provide considerably more information about the relation of the kinetics to the thermodynamics than do many order parameters which are commonly employed by default in the analysis of simulations, such as the radius of gyration or the total number of native contacts (although Q_0 proved to be very useful once trajectories were grouped into appropriate pathways). The essential feature of Q_c and Q_s is that they relax slowly on the time scale of the reaction and thus allow us to average over all the “faster” (less interesting) degrees of freedom.[27,45] If they were ideal reaction coordinates, it would be possible to find a transformation that mapped all possible pairs of Q_c and Q_s to a series of one-dimensional paths, as can be done for reactions of small molecules. However, because of the partial overlap of pathways, such a mapping does not exist. Consequently, the description of the reaction was based on the two-dimensional Q_cQ_s-plane.

5.1. Relation to other Calculations. Several nonstructural coordinates have been suggested for describing protein folding. For example, to describe the behavior of a two-dimensional 13-mer, Chan and Dill defined a set of conformational transitions (moves) for the chain and then determined the minimum number of moves (δ) required to reach the native state.[19,46] Because a state (i) with a value of δ_i cannot reach the native state ($\delta = 0$) without traveling through states with $\delta = \delta_i - 1, \delta_i - 2, ..., 1$, there is a direct correspondence between this coordinate and the ability of the reaction to progress. Unlike the 125-mer, for a small system like the 13-mer, most structures can fold directly to the native state without ever breaking favorable contacts, and the few “off-pathway” structures that do exist typically need to break only one or two favorable contacts to escape. As a result, the energy as a function of δ slopes smoothly towards the native state ($\delta = 0$) [the entropic component of the free energy has a minimum at $\delta = 6$ out of 9, so the the free energy profile (not shown) will depend on temperature].[19,46] Although it provides useful insights into the folding of the 13-mer, the analysis in terms of δ requires enumeration of every possible configuration, so that it cannot be applied to systems of even slightly greater complexity.

Another nonstructural coordinate for describing a reaction is the transmission coefficient (p) itself. Du and co-workers[18,25] have recently suggested that it be used for this purpose. It should be noted that such a type of approach is not fundamentally different from standard activated dynamics simulations[2] like those employed in the present study. Du and co-workers have demonstrated that computers are now sufficiently powerful to perform activated dynamics simulations for a representative sample of the configurations of a simple system (short lattice models), which allows one to skip the usual (structural) analysis of reactive and nonreactive trajectories. For example, they determined the free energy as a function of p for an 18-mer with a sequence that was optimized for fast folding. The transmission coefficient was taken to be the probability that Monte Carlo trials starting from each structure fold before unfolding. The free energy exhibited a small peak at $p = 0.15$ but otherwise had an essentially monotonic downhill slope to the native state ($p = 1$). The transition state, which by definition is at $p = 0.5$, did not correspond to a barrier. The absence of a substantial minimum at $p = 0$ for the 18-mer derives from the fact that there are no low-energy states with $p \approx 0$. In contrast, in the case of the 125-mer, there are many low-energy states which do not interconvert readily to the native state, such as I_1 and I_2. If one could determine the free energy as a function of p_{N_1} (which most closely corresponds to their p in our more complex scheme), the profile would have minima at both $p_{N_1} \approx 1$ (the native basin) and $p_{N_1} \approx 0$ (low-energy traps such as I_1 and I_2).

Recently, an analysis based on p was used to study the kinetics of a 48-mer with a sequence designed for fast folding.[18,47] It was found that the transition states (taken to be those with $0.4 < p < 0.6$) could be grouped structurally into several “classes”, each of which led to a particular “on-pathway” short-lived intermediate. Each class contained several contacts that were common to most or all of the transition states and several contacts that were specific to that class. In our study, we instead grouped transition states by the pathways and intermediates to which they correspond. Each transition state contained with high probability several contacts (for the 125-mer, about 20) that were common to all the transition states and a variable number of contacts that were specific to that transition state, as was the case in ref 47. Our results differ in that the intermediates described for our model are long-lived misfolded (off-pathway) configurations and consequently some of the high probability contacts specific to the transition states leading to those intermediates were nonnative. The differences stem from the fact that Pande and Rokhsar[47] use a Gō interaction set ($B_{ij} = -\epsilon$ for native contacts and $B_{ij} = 0$ for all others)[48] and a much shorter chain length. Although a free energy surface in terms of the number of native contacts (Q_0) and the number of contacts in one of the intermediates (Q_I) is presented, the transition state along the single trajectory shown falls at a free energy minimum in that reduced space. No free energy profile as a function of p is presented, presumably because, unlike Q_c and Q_s, it is not possible to analyze the large amount of equilibrium data that is required with such a computationally costly coordinate.

The free energy as a function of Q_0 for six 125-mer sequences was presented recently.[49] The density of states was determined from simulations that started in the native state and then sampled the space in a manner similar to umbrella sampling except that a series of infinite square well potentials, each of which spanned 11 (out of 176) values of Q_0, was used rather than a series of harmonic potentials. The authors argue that this method should be sufficient to sample adequately the equilibrium phase space because low energy traps with high similarity to the native structure are unlikely to exist. However, such traps are exactly what we found for our system and our attempts to sample by standard umbrella sampling (a harmonic constraint in Q_0) failed in that they converged without ever sampling the “slow tracks”. Due to the sampling method and to the inability of Q_0 by itself to describe the complexities of 125-mer thermodynamics and kinetics, the free energy profiles presented in ref 49 fail to reflect the observed kinetics; for example, at the lowest temperature studied in their model, the chain is reported to become trapped in local minima from which it cannot escape, but no such minima appear in the free energy profile (Figure 7 of ref 49).

The folding mechanism and energy surface presented here for the 125-mer has many elements in common with results obtained from off-lattice models. For example, a simple model of a 46 residue β-barrel has been reported to fold by a nucleation (core formation and sequential growth) mechanism.[50] The mechanism differs somewhat in that at least two hydrophobic clusters of 15−22 residues near the turn regions of the molecule can serve as nuclei, so that the transition state of the off-lattice model varies more in position than does that of the 125-mer (which, as already mentioned, is more variable than that of the lattice 36-mer[37]). However, like the 125-mer, the off-lattice model frequently (but not always) becomes trapped in misfolded intermediates[50] (a “kinetic partitioning mechanism”[51]). The nature of these intermediates has been clarified in a separate study that mapped the topography of the potential surface with a technique previously employed in the analysis of clusters;[52] minima of a subset were enumerated, and the saddlepoints that connect them were determined. The resulting topography of the off-lattice 46-mer has similarities to the 125-mer energy surface. Globally, the 125-mer surface slopes from the open state to the native state (a “staircase topography”, as was found for other “structure-seeking” systems), but there are several low energy minima with structures similar to the native state that are separated from each other and the native state by high barriers.

All-atom simulations are particularly important for addressing issues concerning the roles of side chains and the solvent. One such study recently explored the free energy of folding of a 46 residue three-helix bundle protein from fragment B of staphylococcal protein A.[31,32] The equilibrium sampling of the density of states (in the presence of explicit solvent) was performed by harmonically constraining the chain to particular values of the radius of gyration (R_g); starting configurations were generated by unfolding the chain at high temperature. Given that we sampled only the “fast track” when we did not constrain our lattice system along a coordinate (Q_s) specifically designed to characterize the kinetic intermediates, it is likely that the all-atom simulations do not sample the local minima, if any, that are high in free energy and may play important roles during folding from the denatured state. These simulations are of considerable interest because they make it possible to analyze the thermodynamics with all-atom descriptions. They are likely to describe the free energy surface for the final events of folding (those following formation of a post-critical nucleus); the conformations of importance at that point in the folding reaction are expected to have a relatively native-like topology that would allow the conformational space to be sampled within the timescale accessible to the simulations. The order parameters studied included the radius of gyration, the number of native contacts, and the number of native hydrogen bonds. The “transition state”, which is defined as the states contributing to a barrier of about 3 kcal/mol on the free energy surface as a function of R_g and Q_0, contains about seven (out of 26) native contacts (side chain center-of-mass within 6.5 Å) and has a radius of gyration of 11.7 Å (compared to 9.5 Å for the native state); 50−70% of the native hydrogen bonds are present. Although partially compact, like the transition state of the 125-mer at $T = 0.8$, it appears to have more native-like structure, possibly due to the fact that the native structure of protein A contains primarily helix contacts, which form more readily by chance than do those in sheets.

5.2. Comparison with Experiment. Lattice models are designed to provide an understanding concerning the generic aspects of protein folding. However, it is of interest to determine whether the features found in the simulations correspond to those observed experimentally for specific proteins. Once this connection has been made, it is possible to use the lattice model results to obtain insights into the folding mechanism of those proteins. In so doing as discussed in ref 22, care must be used in translating the results of lattice models to real proteins because of the simplifications used to make the simulations feasible. Each amino acid is represented by a single point, a contact potential is used, and the lattice space is very restrictive. Also, the choice of a local Monte Carlo move set imposes dynamics that do not include large scale diffusive motions.[42−44]

For the 125-mer, folding proceeds along parallel fast and slow pathways. Evidence for a similar kinetic scheme is well documented in lysozyme (for reviews, see refs 13 and 26). This 129-residue protein is composed of an α domain which contains four α-helices, a 3_{10} helix and the chain termini, and of a β domain which contains a triple-stranded β-sheet, a 3_{10} helix and a long loop. The domains are joined by a short, double-stranded antiparallel sheet. Hydrogen exchange labeling in conjunction with two-dimensional NMR has revealed that the amides of the α domain obtain protection within 200 ms while those of the β domain are not fully protected for more than 1 s.[33,53] Electrospray ionization mass spectrometry (ESI-MS) allows decomposition of the averages obtained in NMR into populations of molecules because it provides the distribution of masses within a sample. The formation of a peak in ESI-MS that is intermediate between the unprotected and fully protected forms indicated that lysozyme folds along parallel pathways: a fast one in which the α and β domains fold together and a slow one in which the α domain folds prior to the β domain.[54] This multiple-pathway scheme was confirmed by using fluorescence to monitor interrupted refolding experiments, which are capable of measuring the fraction of native molecules; approximately 14% of molecules follow the fast pathway ($\tau = 50$ ms) and 86% follow the slow pathway ($\tau = 420$ ms).[55] As in the 125-mer, the intermediate appears to be formed by a nucleation−growth mechanism in which the nucleus (core) is intermediate in free energy between the completely unfolded state and the intermediate (in particular, see Figures 4 and 5 of ref 56). However, the dependence of the kinetics on the concentration of denaturant (guanidinium chloride) suggests that lysozyme need not unfold quite as much as the 125-mer to interconvert from an intermediate to the native state.[57]

The intermediates encountered on the slow pathway of the 125-mer have difficulty rearranging to the native state because they must break specific native and nonnative contacts. A nonnative interaction that slows folding has been found in refolding experiments with cytochrome *c*. At pH 6.2, a histidine can replace the native Met 80 ligand of the heme. In the presence of this interaction, the folding populations are heterogeneous and require seconds to reach equilibrium.[58] In contrast, at pHs well below the pK of histidine, where the nonnative contact is not formed, roughly 70% of molecules fold with a 15 ms time constant.[59,60] Thus, as is the case for the 125-mer, nonnative interactions (the histidine-heme contact) can drastically slow folding. Moreover, since the histidines which can form non-native interactions with the heme (26 or 3) are on a loop that does contact the chain termini, it is probable that the docked N- and C-terminal helices, which fold first like the core of the 125-mer, need not come apart to complete the folding process.[15,59] This scenario is similar to that described for the rearrangement of the I_1 state of the 125-mer to its native state at $T = 1.0$; the core remains largely intact while the surface contacts come apart and then form in correct relation to the core.

An example of specific non-native interactions leading to a stable state that does not involve ligation to a metal is provided by bacterial luciferase. Here, folding of the β subunit in the absence of the α subunit leads to a well-structured but enzymically inactive β_2 homodimer instead of the active αβ heterodimer.[61] The similarity in structure between the two subunits (an α-carbon rms deviation of 0.62 Å in the core and 2.6 Å overall) suggests that the homodimer interface is similar to that of the heterodimer, which is composed of a four helix bundle in which each subunit contributes two helices (α2 and α3).[62] To fold to the heterodimer, the homodimer must completely unfold.[61] This behavior is similar to instances in the 125-mer trajectories in which one part of the structure forms in an incorrect orientation relative to another structured part. For example, one of the intermediates found less frequently than I_1 and I_2 had subdomains consisting of residues 1−50 and of residues 51−100 correctly folded, but the two subdomains incorrectly oriented relative to each other. To fold correctly onto the core, residues 1−50 had to unfold completely. For a discussion of other kinetically determined states and their relation to 125-mer folding behavior, see ref 23.

An important question in the study of protein folding is whether intermediates aid or hinder the search for the native state.[63] Although it had been assumed for many years that folding required specific pathways with intermediates, it has now been shown for several proteins[64−67] that folding can proceed rapidly in the absence of intermediates. This has led to the suggestion that all intermediates are "kinetic traps" that hinder folding to the native state.[68] Although the results of the present study and those of the experimental studies described above demonstrate that many intermediates are off-pathway, some folding appears to proceed through mandatory, on-pathway intermediates. One example is barnase (ref 69 and references therein), a 110-residue ribonuclease consisting of a five-stranded β-sheet surrounded by three α-helices; the major hydrophobic core is at the interface between the first α-helix (residues 6−18) and the β-sheet. During folding, this protein populates an intermediate prior to the rate-limiting transition state. Site-directed mutagenesis has been used extensively to characterize the structures of the intermediate and the transition state (the "protein engineering" method). The structure formed in the intermediate, which includes the center of the β-sheet and the last two turns of the first α-helix, is consolidated in the transition state (all the ϕ-values in the transition state are higher than those in the intermediate). This indicates that the intermediate is on-pathway because unfolding need not occur. Moreover, there is no evidence that alternate pathways that lack the intermediate exist (very few fractional ϕ-values). Another example is ubiquitin; although a rapid pre-equilibrium between the unfolded and intermediate states prevents formal kinetic analysis from distinguishing between on- and off-pathway intermediates, mutational effects support a sequential scheme with an on-pathway intermediate.[70,71] No such preequilibrium between the fully denatured state and the intermediates was found in the case of the 125-mer. However, as mentioned above, models with different interactions appear to have such on-pathway intermediates.[18]

The comparison with experimental data illustrates the ability of the simulations. Their goal is not to determine "the" mechanism of protein folding for any specific protein but rather to demonstrate what mechanisms are possible. What makes the lattice approach useful, like any simulation, is that information at levels of detail not accessible to experiment can be obtained. The 125-mer folding reaction is particularly attractive in that it exhibits complexities comparable to those observed for real proteins but it is still simple enough that its kinetics and thermodynamics can be explored fully.

Acknowledgment. We thank Matthias Buck, Themis Lazaridis, and Yaoqi Zhou for helpful discussions. Unless otherwise noted, the plots were generated with the program ASGL by Andrej Šali. A.R.D. is a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported in part by a grant from the National Science Foundation.

Supporting Information Available: Details of the model and the simulations, I_2 transition state analysis, mechanism of core formation, distribution of samples in the reduced coordinates, additional structures, contact maps (native state, intermediates, and transition states), density of states, contour plots of the reaction thermodynamics, equilibrium averages at $T = 1.0$. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Karplus, M.; Porter, R. N.; Sharma, R. D. *J. Chem. Phys.* **1965**, *43*, 3259.

(2) Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959.

(3) Brooks, C. L., III; Karplus, M.; Pettitt, B. M. *Proteins*; John Wiley & Sons: New York, 1988.

(4) Keck, J. *Discuss. Faraday Soc.* **1962**, *33*, 173.

(5) Anderson, J. B. *J. Chem. Phys.* **1973**, *58*, 4684.

(6) Bergsma, J. P.; Reimers, J. R.; Wilson, K. R.; Hynes, J. T. *J. Chem. Phys.* **1986**, *85*, 5625.

(7) Bergsma, J. P.; Gertner, B. J.; Wilson, K. R.; Hynes, J. T. *J. Chem. Phys.* **1987**, *86*, 1356.

(8) Northrup, S. H.; Pear, M. R.; Lee, C.-Y.; McCammon, J. A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 4035.

(9) Neria, E.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 10812.

(10) Fischer, S.; Karplus, M. *Chem. Phys. Lett.* **1992**, *194*, 252.

(11) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902.

(12) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964.

(13) Dobson, C. M.; Karplus, M.; Šali, A. *Angew. Chem. Int. Ed.* **1998**, *37*, 868.

(14) Smith, L. J.; Fiebig, K. M.; Schwalbe, H.; Dobson, C. M. *Folding Des.* **1996**, *1*, R5.

(15) Baldwin, R. L. *Folding Des.* **1996**, *1*, R1.

(16) Karplus, M. Šali, A. *Curr. Opin. Struct. Biol.* **1995**, *5*, 58.

(17) Shakhnovich, E. I. *Curr. Opin. Struct. Biol.* **1997**, *7*, 29.

(18) Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Rokhsar, D. S. *Curr. Opin. Struct. Biol.* **1998**, *8*, 68.

(19) Chan, H. S.; Dill, K. A. *Proteins* **1998**, *30*, 2.

(20) Šali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.

(21) Karplus, M. In *Simplicity and Complexity in Proteins and Nucleic Acids*; Frauenfelder, H., et al., Eds.; Dahlem University Press: Berlin, 1999.

(22) Dinner, A. R.; Šali, A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8356.

(23) Dinner, A. R.; Karplus, M. *Nature Struct. Biol.* **1998**, *5*, 236.

(24) Dinner, A. R.; So, S.-S.; Karplus, M. *Proteins* **1998**, *33*, 177.

(25) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334.

(26) Dobson, C. M.; Evans, P. A.; Radford, S. E. *Trends Biochem. Sci.* **1994**, *19*, 31.

(27) Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press: New York, 1987.

(28) Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1990**, *92*, 3118.

(29) Frantz, D. D.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **1990**, *93*, 2769.

(30) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.

(31) Boczko, E. M.; Brooks, C. L., III *Science* **1996**, *269*, 393.

(32) Guo, Z.; Brooks, C. L., III; Boczko, E. M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10161.

(33) Radford, S. E.; Dobson, C. M.; Evans, P. A. *Nature* **1992**, *358*, 302.

(34) Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. *Biochemistry* **1995**, *34*, 3066.

(35) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A.-M. *Proteins* **1995**, *23*, 142.

(36) Dinner, A. R.; Karplus, M. *J. Mol. Biol.* **1999**. In press.

(37) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Biochemistry* **1994**, *33*, 10026.

(38) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *J. Chem. Phys.* **1994**, *101*, 6052.

(39) Mirny, L. A.; Abkevich, V.; Shakhnovich, E. I. *Folding Des.* **1996**, *1*, 103.

(40) Dinner, A. R.; Abkevich, V.; Shakhnovich, E.; Karplus, M. *Proteins* **1999**, *35*, 34.

(41) Hammond, G. S. *J. Am. Chem. Soc.* **1955**, *77*, 334.

(42) Karplus, M.; Weaver, D. L. *Nature* **1976**, *260*, 404.

(43) Karplus, M.; Weaver, D. L. *Biopolymers* **1979**, *18*, 1421.

(44) Karplus, M.; Weaver, D. L. *Prot. Sci.* **1994**, *3*, 650.

(45) Du and co-workers[25] term the degree(s) of freedom which relax most slowly "transition coordinate(s)" and argue that the term "reaction coordinates" should be reserved for the lengths along specific (one-dimensional) paths in the complete (many-dimensional) coordinate space (such as are obtained in using automatic methods[10,12]). We use the term "reaction coordinates" to describe their "transition coordinates" because we feel that a generalized notion of reaction coordinates is already prevalent in the literature of complex systems (for example, see ref 8).

(46) Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1994**, *100*, 9238.

(47) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1273.

(48) Taketomi, H.; Ueda, Y.; Gō, N. *Int. J. Peptide Protein Res.* **1975**, *7*, 445.

(49) Chung, M. S.; Neuwald, A. F.; Wilbur, W. J. *Folding Des.* **1997**, *3*, 51.

(50) Guo, Z.; Thirumalai, D. *Folding Des.* **1997**, *2*, 377.

(51) Thirumalai, D.; Woodson, S. A. *Acc. Chem. Res.* **1996**, *29*, 433.

(52) Berry, R. S.; Elmaci, N.; Rose, J. P.; Vekhter, B. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 9520.

(53) Miranker, A.; Radford, S. E.; Karplus, M.; Dobson, C. M. *Nature* **1991**, *349*, 633.

(54) Miranker, A.; Robinson, C. V.; Radford, S. E.; Aplin, R. T.; Dobson, C. M. *Science* **1993**, *262*, 896.

(55) Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9029.

(56) Kiefhaber, T.; Bachmann, A.; Wildegger, G.; Wagner, C. *Biochemistry* **1997**, *36*, 5108.

(57) Wildegger, G.; Kiefhaber, T. *J. Mol. Biol.* **1997**, *270*, 294.

(58) Roder, H.; Elöve, G. A.; Englander, W. S. *Nature* **1988**, *335*, 700.

(59) Elöve, G. A.; Bhuyan, A. K.; Roder, H. *Biochemistry* **1994**, *33*, 6925.

(60) Sosnick, T. R.; Mayne, L.; Hiller, R.; Englander, S. W. *Nature Struct. Biol.* **1994**, *1*, 149.

(61) Sinclair, J. F.; Ziegler, M. M.; Baldwin, T. O. *Nature Struct. Biol.* **1994**, *1*, 320.

(62) Fisher, A. J.; Raushel, F. M.; Baldwin, T. O.; Rayment, I. *Biochemistry* **1995**, *33*, 6581.

(63) Laurents, D. V.; Baldwin, R. L. *Biophys. J.* **1998**, *75*, 428.

(64) Jackson, S. E.; Fersht, A. R. *Biochemistry* **1991**, *30*, 10428.

(65) Huang, G. S.; Oas, T. G. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6878.

(66) Schindler, T.; Herrier, M.; Marahiel, M. A.; Schmid, F. X. *Nat. Struct. Biol.* **1995**, *2*, 663.

(67) Eaton, W. A.; Muñoz, V.; Thompson, P. A.; Chan, C.-K.; Hofrichter, J. *Curr. Opin. Struct. Biol.* **1997**, *7*, 10.

(68) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *Proteins* **1998**, *32*, 136.

(69) Fersht, A. R. *FEBS Lett.* **1993**, *325*, 5.

(70) Khorasanizadeh, S.; Peters, I. D.; Roder, H. *Nature Struct. Biol.* **1996**, *3*, 193.

(71) Roder, H.; Colón, W. *Curr. Opin. Struct. Biol.* **1997**, *7*, 15.

(72) Nelson, M.; Humphrey, W.; Gursoy, A.; Dalke, A.; Kalé, L.; Skeel, R.; Schulten, K.; Kufrin, R. *Comput. Phys. Commun.* **1995**, *91*, 111.