Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 8356–8361, August 1996
Biophysics

The folding mechanism of larger model proteins: Role of native structure

(protein folding/lattice model/Monte Carlo/secondary structure/folding bottlenecks)

AARON R. DINNER*†, ANDREJ ŠALI†‡, AND MARTIN KARPLUS*†§

*Committee on Higher Degrees in Biophysics and †Department of Chemistry, Harvard University, Cambridge, MA 02138; ‡The Rockefeller University, 1230 York Avenue, New York, NY 10021; and §Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, 67000 Strasbourg, France

Contributed by Martin Karplus, May 2, 1996

ABSTRACT **The folding mechanism of a 125-bead heteropolymer model for proteins is investigated with Monte Carlo simulations on a cubic lattice. Sequences that do and do not fold in a reasonable time are compared. The overall folding behavior is found to be more complex than that of models for smaller proteins. Folding begins with a rapid collapse followed by a slow search through the semi-compact globule for a sequence-dependent stable core with about 30 out of 176 native contacts which serves as the transition state for folding to a near-native structure. Efficient search for the core is dependent on structural features of the native state. Sequences that fold have large amounts of stable, cooperative structure that is accessible through short-range initiation sites, such as those in anti-parallel sheets connected by turns. Before folding is completed, the system can encounter a second bottleneck, involving the condensation and rearrangement of surface residues. Overly stable local structure of the surface residues slows this stage of the folding process. The relation of the results from the 125-mer model studies to the folding of real proteins is discussed.**

A protein molecule is composed of a polypeptide sequence that meets both thermodynamic and kinetic folding requirements: it possesses a unique native state that is stable at a physiological temperature, and it is able to find that state in a reasonable time (milliseconds to minutes) at that temperature in spite of the fact that the number of possible configurations for the polypeptide prevents it from making an exhaustive search (Levinthal paradox) (1). Many Monte Carlo (MC) folding studies for short chaius have been reported (see, for example, refs. 2–5). Simulations of 27-mer random heteropolymer sequences on a cubic lattice showed that a necessary and sufficient condition for satisfying the folding requirements of that model is that the native state is a pronounced global energy minimum with a large energy gap between the ground state and the rest of the states. There is no correlation between the rate of folding and any other sequence attributes, such as the amount of secondary structure in the native state. Folding proceeds by a fast ($\approx 10^4$ MC steps) collapse to a semi-compact random globule, followed by a slow ($\approx 10^7$ MC steps), nondirected search through the $\approx 10^{10}$ semi-compact structures for one of the $\approx 10^3$ transition states that lead rapidly (within 10^5 MC steps) to the native conformation (5). The restriction of the search to the compact portion of the conformation space and the large number of transition states lead to the resolution of the Levinthal paradox for the 27-mer. The energy gap criterion results in a native state that is stable at a temperature high enough for the folding polypeptide chain to overcome barriers between random semi-compact states.

The 27-mer mechanism appears to be limited to small proteins (5) because the time to fold increases exponentially with the length of the chain and becomes unrealistically long for more than about 80 residues. To determine what factors could be involved in resolving the Levinthal paradox for longer chains, we consider 125-mers with $5 \times 5 \times 5$ fully compact native states, which have a length and a surface-area-to-volume ratio (22%) corresponding to that found in globular proteins (6). Conceptually, such 125-mers can be regarded as a 27-mer $3 \times 3 \times 3$ cube embedded in a layer of surface monomers. To investigate the role of native structure in the folding process, we engineer sequences with high secondary structure content in the native state. The introduction of nonrandom properties to facilitate folding does not bias the results, since, as in the 27-mer study (2), we base our conclusions on comparisons of similarly engineered sequences with varying degrees of folding abilities.

METHODS

We generated a data base of 200 sequences, tested each for its ability to fold, and looked for sequence attributes that have a high correlation with rapid folding. Such an approach requires that the sequences that make up the data base have a range of folding abilities. The probability of finding random sequences that satisfy both the thermodynamic and kinetic requirements for folding is very small for a 125-mer. Unlike the 27-mer case, where 54 out of 200 sequences were found to fold (2), none was found in ten 125-mer sequences. If the fraction of sequences satisfying the thermodynamic requirement is large ($\approx 1\%$) and independent of chain length, as proposed by Shakhnovich and Gutin (7), this suggests that the stability criterion is not sufficient for satisfying the kinetic requirement for folding in the 125-mer. Consequently, to obtain a 125-mer data base with a significant fraction of folding sequences, nonrandom properties must be introduced. To examine the possible role of secondary structure (see Table 1 and ref. 8), we designed sequences with substantial amounts of helices and sheets. From an ensemble of fully compact $5 \times 5 \times 5$ (176-contact) conformations, we selected 100 otherwise random structures with a large fraction (between 40 and 100%) of contacts in sheets and 100 otherwise random structures with a large fraction (between 34 and 54%) of contacts in helices. Each of the 200 chosen structures serves as the starting point for a design process that creates a sequence for which that structure is the native (lowest energy) state. The resulting sequences can be regarded as representing, respectively, the β-sheet and α-helical classes of proteins (9).

The energy function used here has the same form as that employed previously (2, 5). It is the sum over all contacts: $E = \sum_{i<j} \Delta(r_i - r_j)(B_{ij} + B_0)$. The function $\Delta(r_i - r_j)$ selects the interacting monomer pairs; it is unity if i and j, located at r_i and r_j, respectively, are in contact and zero otherwise. B_{ij} gives the specific interaction energy between monomers i and j; a complete set of B_{ij} defines a sequence. The quantity B_0 ($B_0 < 0$) is a

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: MC, Monte Carlo; T_m, melting temperature.

mean attraction between monomers that corresponds to an overall hydrophobic term.

To ensure that the chosen structures had the lowest energy, the B_{ij} values are selected from two Gaussian distributions: native B_{ij} (B_{ij}^{ε}) with a mean of $\varepsilon = -1.67$ and standard deviation $\sigma_B \approx 1.0$ and nonnative B_{ij} (B_{ij}^{0}) with $\varepsilon = 0$ and $\sigma_B \approx 1.0$. Since there are far more nonnative contacts (3606 compared to 176), the overall B_{ij} distribution has essentially the same mean and standard deviation as the B_{ij}^{0} random distribution, which corresponds to that used in the 27-mer studies (2, 5). For 27-mer sequences generated in this manner, 196 out of 200 have a large energy gap and so met the folding requirements of ref. 2. On this basis, we expect and find that most 125-mer optimized sequences meet the thermodynamic requirement, which is expected to lead to an increase in the fraction of sequences capable of folding.

The 200 sequences were each subjected to 10 independent Metropolis MC trials (10), each of which started with a different random configuration. Trials continued for a maximum of 50×10^6 MC steps and were terminated earlier if the chain found the native state (first passage time). The allowed chain moves are tail flips, internal bead flips, crankshaft rotations, and a composite move in which two to four single monomer or crankshaft moves are performed before applying the Metropolis criterion. The fraction of composite moves (distributed equally between 2, 3, and 4 simple moves) was 0.4, and the fraction of crankshaft moves (including those that are part of a composite move) was 0.4. Based on preliminary trials, B_0 was set to -0.60 and the temperature factor (k_BT) was 1.34 for all trials; the latter corresponds to a range that yields rapid folding for the 27-mer sequences (11). MC sampling of the entire conformational space shows that this temperature approximates T_m for most of the sequences—i.e., they have a distribution of native state probabilities roughly centered on 0.50 with a standard deviation of 0.21.

As we show below, the folding behavior of the 125-mer sequences is more complex than that observed in previous lattice simulations. Consequently, the analysis cannot be based simply on sequences that repeatedly fold (reach the native state) and those that do not. Instead, the folding behavior of a sequence is measured both by the number of times it reached the native state in 10 trials (N_f) and by the average contact overlap between the native structure (Q_0) and the lowest energy structure sampled during each trial (Q_m). Although Q_m is not the highest Q_0 (computation of the latter requires more time), Q_m was typically observed to be within 0.05 of the highest Q_0 in a total of 40 trials with 20 sequences; the Pearson correlation coefficient between these two quantities was 0.99. Q_m identifies sequences that get close to the native state but fail to find it.

RESULTS

The energy gap is determined by the scaled energy separation of the native and a random maximally compact state (12–14): it is given by $\Delta/\sigma_B = [176(B_0 + \varepsilon) - E_0]/\sigma_B$ where ε and σ_B are the mean and standard deviation of the overall B_{ij} distribution, and 176 is the number of contacts in the maximally compact conformation. Although there is only a weak correlation between N_f and Δ/σ_B, there is a strong correlation between Q_m and Δ/σ_B (Table 1). This shows that the energy gap criterion is important for determining whether the protein gets near the native state ($Q_m > 0.8$), which is a necessary but not sufficient condition for 125-mer folding. This result contrasts with those for the 27-mer for which such a large energy gap is necessary and sufficient for complete folding (2, 5), as it is in designed 27-mers to 80-mers (12–14).

To examine the folding from a random coil to a near-native structure (native contact overlap $Q_0 > 0.8$), we plot the average first passage time for a given native contact overlap

Table 1. Comparison of sequence attributes and folding

	All		Helix		Sheet		$Q_m > 0.8$
	N_f	Q_m	N_f	Q_m	N_f	Q_m	N_f
Δ/σ_B	0.24	0.48	0.29	0.61	0.22	0.42	−0.20
$C_0(3)$	−0.48	−0.46	0.16	0.06	0.25	0.26	−0.33
$C_0(5)$	−0.40	−0.33	−0.11	0.04	0.32	0.34	−0.31
$C_0(7)$	0.51	0.47	0.08	0.09	0.31	0.28	0.44
$C_0(9)$	0.44	0.41	−0.11	−0.08	0.17	0.19	0.31
C_h	−0.52	−0.47	0.02	0.17	−0.12	−0.07	−0.39
C_{ps}	−0.21	−0.16	−0.06	−0.02	−0.33	−0.32	−0.23
C_{as}	0.55	0.50	0.08	0.14	0.40	0.36	0.44
C_t	0.56	0.50	0.19	0.13	0.38	0.35	0.47

Pearson linear correlation coefficients (r_{xy}) between sequence attributes (x) and folding ability measures (y): $r_{xy} = \sigma_{xy}^2/\sigma_x\sigma_y = \Sigma_i^n(x_i - \langle x\rangle)(y_i - \langle y\rangle)/\sqrt{[\Sigma_i^n(x_i - \langle x\rangle)^2\Sigma_i^n(y_i - \langle y\rangle)^2]}$. A perfect correlation has $r_{xy} = 1$, a perfect anti-correlation has $r_{xy} = -1$. For the present sample sizes (see text), $|r_{xy}| \geq 0.27$ provides at least 99% certainty of significance and $|r_{xy}| \geq 0.45$ is enough to yield predictive accuracy. Δ/σ_B is defined in the text. $C_0(k)$ is the number of contacts in the native structure with contact order $k = |i - j|$, and C_h, C_{ps}, C_{as}, and C_t are the numbers of contacts in helices, parallel sheets, anti-parallel sheets, and turns, respectively. All elements of secondary structure must have at least two contacts to be counted; helices must have at least one complete turn. A contact may belong to more than one secondary structure group. See ref. 8 for full definitions.

[$\langle t(Q_0)\rangle$] for two sequences, one sheet and one helical (Fig. 1). There is a rapid collapse ($\approx 10^6$ MC steps) to a compact globule with about 120 contacts (out of 176 possible) but little native structure ($Q_0 \approx 0.25$, corresponding to random overlap). There is then a slow search ($\approx 20 \times 10^6$ MC steps) among semi-compact states for a particular set of about 30 contacts that form a spatially localized core in the native structure (Fig. 2). At the time the core is formed, the chain typically has about 60 native contacts ($Q_0 \approx 0.35$): 30 in the core and 30 in the rest of the chain due to random overlap with the native state. Formation of this core occurs by a mechanism similar to the complete folding of a 27-mer—i.e., there is a random search for any one of a relatively large number of transition states that lead rapidly to the core structure. The transition states have a pairwise core contact overlap of about 70% with each other. This is slightly lower than that of the 27-mer transition states (85%) (15). Although 27-mers often globally unfold after finding the native state for the first time (5), the core does not, and the polymer chain rapidly condenses around it until $Q_0 \approx 0.8$ is reached. The nuclear role of the core is confirmed by the second set of curves in Fig. 1, which show rapid folding for trials in which the core contacts are introduced (see Fig. 1 legend for procedure).

For folding to high Q_0 by this mechanism, the core must be readily found among the large number of semi-compact states, and it must be stable enough not to unfold before most of the chain condenses around it. The core contacts have average B_{ij} of -2.00, compared to the overall average native B_{ij} of -1.67. The range of the core B_{ij} values is between about -3.35 and 0.85, so that they are not simply the lowest energy native contacts. Nevertheless, the overall low energy of the core contacts and the correlations between Q_m and Δ/σ_B indicate that the effectiveness of the core is directly related to its stability. This result was expected from the similarity between core formation and the folding of a 27-mer or a 36-mer; in the latter cases, increase in the stability of native contacts relative to others was shown to increase the energy gap and the folding rate (12, 14).

In contrast to the random 27-mer folding sequences, secondary structure plays an essential role in reducing the time required to search through the semi-compact globule for the core of the 125-mer. For the 27-mer, there is a rapid interconversion between globally different semi-compact structures

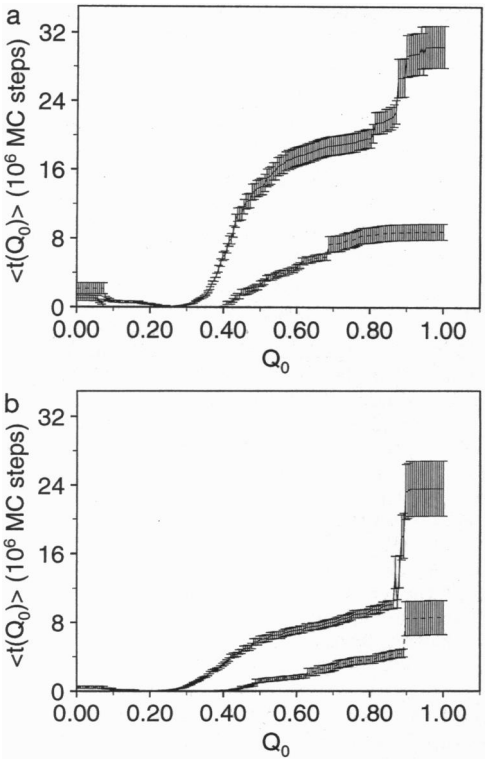

FIG. 1. Average first step [$\langle t(Q_0)\rangle$] at which a given fractional contact overlap with the native state (Q_0) is reached for (*a*) a helix sequence with $N_f = 8$ and $Q_m = 0.90341$ and (*b*) a sheet sequence with $N_f = 10$ and $Q_m = 1.0$. Solid lines are for trials beginning in random configurations; dashed lines are for trials in which the core contacts are introduced, the other contacts are randomized at high temperature ($T = 16{,}667$), and folding is subsequently allowed to proceed without any constraints at the normal folding temperature ($T = 1.34$). Each average is over 50 trials; the error bars are the standard error of the mean. 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.

that allows a full random search for a transition state in the relatively small configuration space ($\approx 10^{10}$ structures and $\approx 10^3$ transition states). Such a search is not possible for the 125-mer where global interconversion is much slower and the space is much larger ($\approx 10^{47}$ structures and $\approx 10^{14}$ transition states). Analysis shows that it is important for core formation that a sequence have cooperative structures in its native state. Contacts are "cooperative" if formation of any one contact increases the probability of formation of the others. Further, the presence of short-range contacts within the cooperative structures allows them to be found more readily by a nondirected search. The structures which exhibit these characteristics most clearly are anti-parallel sheets and turns; Table 1 shows that there is a high positive correlation between Q_m and both of these attributes ($r_{C_t Q_m} = 0.50$ and $r_{C_{as} Q_m} = 0.50$; see Table 1 for definitions). The importance of cooperativity becomes apparent from comparison of these correlations to anti-correlations between Q_m and helical contacts ($r_{C_h Q_m} = -0.47$). Since both turns and helices involve local contacts and differ only in whether they are followed by intermediate- or long-range cooperative contacts, the difference demonstrates the importance of cooperativity in folding to high Q_0. The role of readily found short-range contacts is illustrated by the anti-correlation between Q_m and the cooperative parallel sheet contacts ($r_{C_{ps} Q_m} = -0.16$). Parallel sheets have the same cooperative features as anti-parallel sheets; the difference between the two is that the former lack the short-range turn contacts that serve as initiation sites in the latter. The correlations between Q_m and secondary structure are in accord with the correlations with $C_0(k = |i - j|)$ (Table 1). For $C_0(3)$ and $C_0(5)$, the correlation is negative when considering all sequences since helical sequences, which fold less often, have many more $k = 3$ and $k = 5$ contacts. However, there is a weak positive correlation when one looks only at sheet sequences where most of the $k = 3$ and $k = 5$ contacts are in turns. The $k = 7$ and $k = 9$ contacts are found most frequently in anti-parallel sheets and consequently exhibit high correlations with Q_m; interestingly, $r_{C_{as} Q_m} > r_{C_0(7) Q_m} > r_{C_0(9) Q_m}$, showing that having more cooperative contacts is more important than having more intermediate- or long-ranged ones, and that simply maximizing the number of longer-ranged contacts does not necessarily optimize the folding rate.

The energy gap and secondary structure, represented by Δ/σ_B and C_t in Fig. 3, can be used together to distinguish sequences which repeatedly fold to high Q_0 from those that do not; a corresponding diagram is obtained with Δ/σ_B and C_{as}. That both the stability and structural features are important is confirmed by the fact that the line drawn to maximize the number of correct predictions is not vertical. This partial uncoupling of stability and folding by the introduction of secondary structural elements goes beyond the 27-mer results and suggests that more cooperative structure can compensate for less energetic optimization.

Once $Q_0 > 0.8$ is reached, many sequences encounter a second kinetic bottleneck, as indicated by the jump in $\langle t(Q_0)\rangle$ near the native state (Fig. 1). This final stage of folding involves the rearrangement and condensation of surface residues (Fig. 2). When the chain first reaches high Q_0, these residues are typically in a misfolded local globular structure that is separate from the core. Such local structures must dissolve before correct condensation can take place. During this stage of the simulations, the chain may unfold to as low as $Q_0 = 0.5$ prior to rapidly refolding. These fluctuations are not equivalent to the global unfolding observed for a collapsed 27-mer since the core remains folded.

To understand the final stage of folding, we examine certain statistical correlations. For the 137 sequences with $Q_m > 0.8$, there is an anti-correlation between Δ/σ_B and N_f ($r_{\Delta/\sigma_B N_f} = -0.20$; see Table 1), which indicates that better folding sequences tend to have contacts with less individual stability in noncore regions. This is confirmed by the fact that the anti-correlation is even stronger (-0.28) if one looks only at surface contacts. The contrast with the high positive correlation for core formation demonstrates that overstabilization makes it difficult for the chain to convert misfolded surface regions that occur in the simulations into the native structure. This slows down completion of folding. The correlations between cooperative structure folding ($r_{C_{as} N_f}$ and $r_{C_t N_f}$) for the 137 sequences are lower but in the same direction as those for core formation—i.e., cooperative structure accelerates the final stage of folding, but less so than for core formation. Sequences with more cooperative native structure have a lower random overlap with misfolded structures, but interconvert between near native structures more slowly. This suggests that rapid completion of the folding reaction involves a balance between simplification of the search by the presence of cooperative elements and overstabilization of misfolded structures. At low temperature, a free energy barrier is present that is reduced as the temperature is raised. This is in accord with the primarily energetic origin of the final folding bottleneck. The most efficient folding sequences are ones that avoid getting trapped in structures which have local similarity to the native state but lack the overall fold. Similar behavior was observed in one of the pioneering lattice simulations (16) in which the only attractive interactions were between monomers in contact in the native state.

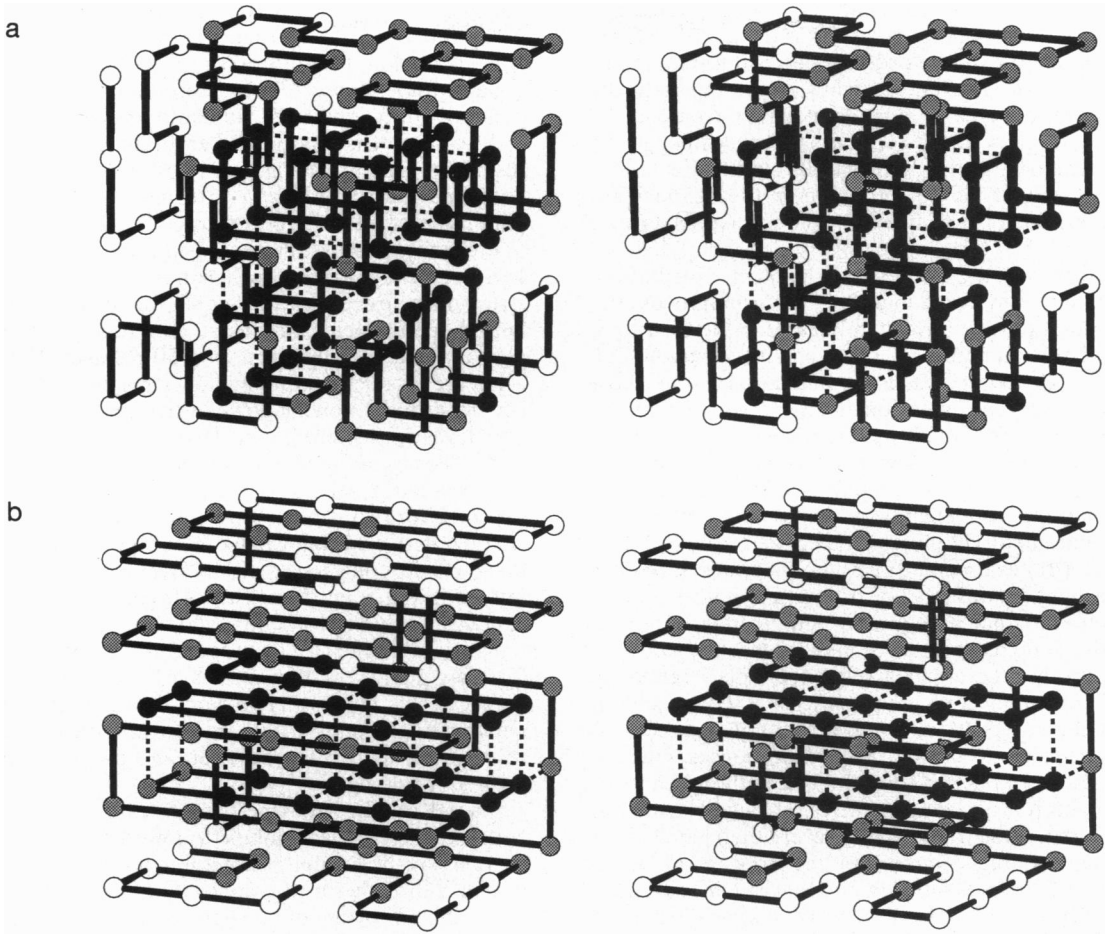

FIG. 2. Stereodiagrams of the relative order of folding mapped on the native structure for the sequences shown in Fig. 1. The contacts of core monomers are shown by dashed lines. Black monomers have probability ≥75% of being in their native position relative to the core in the last structures at $Q_0 = 0.5$, white monomers have probability ≤75% of being in the native position relative to the core in the last structures at $Q_0 = 0.80$, and gray monomers include all others. Probabilities are based on 50 trials.

DISCUSSION

A comparison of folding and nonfolding 125-mer sequences with either helical or sheet native structure was made to distinguish the characteristics necessary for folding from those common to all such sequences that are thermodynamically stable. This comparison is an essential element of the present approach and previous 27-mer studies (2, 5) that differentiates them from experiments and most other folding simulations, which consider only sequences that fold (but see refs. 17 and 18). A large number of sequences (200) were studied to make possible meaningful generalizations.

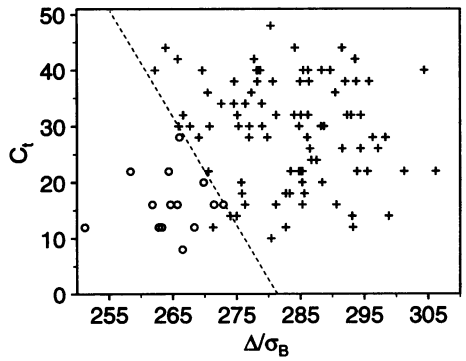

FIG. 3. Folding predictions with a combination of stability and structural measures. Comparison of the (+) 92 sequences with $Q_m > 0.9$ to the (○) 14 with $Q_m < 0.5$. The correlation coefficient between Δ/σ_B and C_t is 0.038, indicating independence. The dotted line is drawn to separate folding and nonfolding sequences [its equation is $C_t = -1.92(\Delta/\sigma_B) + 541.15$].

Complexities not evident in simulations of shorter chains appear in the folding of the 125-mers. The folding mechanism consists of two separate stages. The first stage involves formation of a core that serves as a nucleus for folding to a near-native (≈80%) structure, and the second stage involves rearrangement and condensation of the remaining residues around the core to form the native structure. The cores consist of about 30 monomers with ≈30 native contacts and have a compact but not entirely regular form. A different core is associated with each sequence, each of which has a different native state.

During the folding process, the core is found by an extension of the 27-mer random search mechanism. Formation of the core is facilitated by both the stabilities of its contacts and the presence of cooperative secondary structure with effective initiation sites. The strong dependence of the ability of a sequence to fold in a reasonable time on both the distribution and position of its native contacts contrasts sharply with the random 27-mer chains (2). The essential difference is that the 27-mer collapsed globule can rearrange rapidly enough to do an efficient random search; in the 125-mer, the need for such a complete rearrangement is obviated by the formation and persistence of cooperative structure that restricts the search.

The results for formation of the core in the 125-mer are different from the nucleation mechanism found for folding 36-mer sequences that were optimized to have a large energy

gap without any structural considerations (19). In that case, about eight obligatory monomers served as the nucleus. Thus, for chains with lengths on the order of 36 residues, there are at least two different folding mechanisms for energetically optimized sequences. One is a random search restricted by secondary structure, as in the 125-mer, and the other is the nucleation mechanism of the 36-mer (19). Unlike the 36-mer nucleus, few, if any, of the core contacts in the 125-mer are absolutely necessary for it to function as a transition state to higher Q_0.

The importance of the native state contact distribution, which was found to play a role in both core formation and the final condensation to the native state, has been the focus of a number of lattice simulations of short chains. Govindarajan and Goldstein (20) used exhaustive enumeration of short chains in two and three dimensions to demonstrate that the structures which are more readily energetically optimizable by their criterion are those with fewer short-range interactions. Since they did not study the actual folding kinetics, the relevance of their results depends on the assumption that the optimization parameter is closely linked to folding ability. Abkevich *et al.* (21) examined three highly optimized 36-mer sequences, each with a different distribution of native contact orders. They found that a sequence with a (nonmaximally compact) native state having a large number of $k = 3$ contacts folded more slowly than either a sequence with a randomly chosen $3 \times 3 \times 4$ cube as its native state or a sequence with a (nonmaximally compact) native state without any $k = 3$ contacts. Their observation that the two sequences with cooperative contacts fold better than the one with only noncooperative contacts is in agreement with the strongest structural correlation of the present study. However, their conclusion that short-range contacts hinder folding appears to be in disagreement with the present work and that of Dill *et al.* (22). To understand the origin of this difference, it should be noted that the 36-mer of ref. 21 with primarily short-range contacts is noncompact and has almost no long-range interactions; it lacks the cooperative contacts that we found to be essential for fast folding to a near-native structure. It corresponds more closely to the model introduced by Zwanzig *et al.* (23) in which the local biases toward the native conformation are more important than the tertiary interactions. Folding of such a model is noncooperative and the chain spends a significant fraction of its time in conformations intermediate between the completely folded and unfolded states (24). The 125-mer helical sequences have on the order of half their interactions in long-range contacts and exhibit the cooperative "all-or-none" transition characteristic of protein folding. In addition, the greater length of the 125-mer sequences permits a more coarse-grained analysis of short- versus long-range contacts. For example, there is little difference in the time it takes to find a $k = 3$ and a $k = 5$ contact by a random search, so we consider both to be short range. From this point of view, the other two 36-mer sequences of ref. 21, one of which has no $k = 3$ contacts and one of which is "random," are in the same "class" (a class which is more similar to sheet sequences than to helical ones). Also, the small size of the 36-mer reduces the importance of readily accessible initiation sites for cooperative structures because they cannot have really long-range contacts, which puts fewer kinetic accessibility demands on the chain.

Dill *et al.* (25) addressed the issue of accessibility and ease of propagation in the "hydrophobic zipper" hypothesis. In this scheme, short-range hydrophobic contacts form first and facilitate the subsequent formation of cooperative long-range hydrophobic contacts by restricting the chain. In disagreement with one aspect of the hydrophobic zipper hypothesis, native secondary structure formation follows collapse to a compact disordered globule in the present study. This "burst collapse" is a consequence of the choice of B_0 and was studied in ref. 26. Although compactization may slow certain steps in the folding mechanism by making rearrangement more difficult, it speeds up a random search by reducing the number of states. As for core formation, the different models show that both random collapse or collapse to an ordered structure can occur, depending on the choice of simulation parameters.

A variety of comparisons can be made with experimental results for protein folding. The role of initiation sites and stable cooperative structure is in agreement with the statistics for data bases of known protein structures. Brocchieri and Karlin (27) classify spatially adjacent residues by sequence separation. In agreement with the conclusion that long-range cooperative structure helps restrict the search for the core, they find that buried amino acids are most commonly separated in the primary sequence by more than 50 residues. Exposed amino acids are most commonly within 4 residues of each other, in accord with the slowing down of the final stages of folding for structures with more cooperative, longer-range surface contacts. The largest number of strand–strand contacts were those with sequence separations between 5 and 20 residues, as expected for two strands separated by a tight turn. Rooman *et al.* (28) derived a knowledge-based potential of mean force that considers only local (within 8 residues along the sequence) interactions to identify structural fragments (between 5 and 15 residues) with large stability; where data were available, such units were found to correspond to early folding regions. Corresponding initiation sites have been studied by Avbelj and Moult (29). These results and solution measurements of peptide fragments (30) are in accord with the lattice model results that tight turns between anti-parallel sheets tend to favor fast folding.

Studies of hen lysozyme demonstrate that there is a rapid collapse to a compact globule followed by rearrangement to a more native-like state (31). Although human lysozyme differs in that it directly forms a correctly folded stable and cooperative core consisting of two helices at the N terminus and one at the C terminus (32), both sequences exhibit kinetics indicating that there exist parallel fast and slow pathways, in analogy to the 125-mer.

For cytochrome *c*, folding studies indicate a "burst" collapse, followed by a docking of the N- and C-terminal helices (33). This early structural feature corresponds to the most stable of a series of metastable partially unfolded forms, each of which differs from the previous one by the sequential loss of a cooperative folding unit (34). Thus, folding is thought to proceed by formation of a stable core (the pair of terminal helices and their hydrophobic contacts) followed by sequential condensation of the rest of the protein, in analogy to the result of the 125-mer. Cooperativity in the folding units thus appears to be essential for both the thermodynamics and kinetics of cytochrome *c* folding.

Core formation followed by sequential growth was observed in mutational analyses of folding of both CI2 (35) and barnase (36). Although one or two residues appear to be involved in essential contacts, most are present only part of the time in the transition state for folding, as for the 125-mer. CI2 exhibits two-state kinetics and lacks an observable barrier near the native state (35). Although some of the 125-mer sequences also fold immediately from the near native to the native state, many exhibit more complex behavior similar to barnase (35, 36) and lysozyme (31, 32), which have transition states near the end of the folding pathway.

Toward the end of the folding process, the residues at the surface are found to rearrange and condense to form the native state. Although the chain does not undergo global rearrangement, as in the 27-mer, the monomers added in the sequential growth stage often rapidly unfold and refold, reaching Q_0 values as low as 0.5. Partial core restructuring occurs in some folding trajectories, but the dominant role is played by the surface residues. The folding kinetics of the bovine pancreatic inhibitor involve a slow step that disrupts native structure, even

in the core, so as to be able to rearrange disulfide bonds (37, 38).

Although the above comparisons with real proteins are important, care is needed in applying the results of lattice models. First, the models are highly simplified and use an effective nearest-neighbor potential. There is no explicit consideration of the various types of interactions (e.g., hydrophobic, polar, hydrogen bonding) or the role of side chains. Although the successes in using similar potentials for protein "threading" suggest that this may not be an essential limitation (39, 40), particularly for general aspects of the folding problem, it is a factor to consider when making comparisons with specific proteins. Second, the MC move set imposes a certain dynamics on the folding behavior of the chains. For example, there are no large scale motions of semi-rigid units like helices which eliminates mechanisms of the diffusion-collision type (41). In spite of this limitation, the present model does not yield unrealistically long folding times; with a step time of 10^{-9} s, corresponding to a C_α pseudo-dihedral angle transition, 20×10^6 MC steps yield a 20 ms folding time.

The most important role of the simulations is not to determine "the" mechanism of protein folding (there are likely to be several when finer details are considered), but rather to demonstrate what mechanisms are possible. This will hopefully stimulate experiments to show which, if any, of the suggested mechanisms operates in a certain protein. What makes the lattice approach useful, like any simulation, is that information at all levels of detail can be determined and that the parameters can be varied easily to demonstrate their role in the observed mechanism.

Note Added in Proof. The recent paper by Sosnick *et al.* (42) gives more details of the cytochrome *c* folding mechanism that is closely related to the 125-mer results.

We thank E. Shakhnovich, A. Gutin, and V. Abkevich for providing some optimized sequences for comparison and C. M. Dobson and W. Eaton for their comments on the manuscript. Simulations were done on an Iris 4D, an IBM RS6000-550, DEC Alphas, a DECstation 5000/240, and HP 735. This work was supported in part by a grant from the National Science Foundation. A.R.D. is a Howard Hughes Medical Institute Pre-Doctoral Fellow. A.Š. was a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research.

1. Levinthal, C. (1969) in *Mossbauer Spectroscopy in Biological Systems*, ed. Debrunner, P. (Univ. of Illinois Press, Urbana).
2. Šali, A., Shakhnovich, E. & Karplus, M. (1994) *J. Mol. Biol.* **235,** 1614–1636.
3. Chan, H. S. & Dill, K. A. (1994) *J. Chem. Phys.* **100,** 9238–9257.
4. Socci, N. D. & Onuchic, J. N. (1994) *J. Chem. Phys.* **101,** 1519–1528.
5. Šali, A., Shakhnovich, E. & Karplus, M. (1994) *Nature (London)* **369,** 248–251.
6. Richards, F. M. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), pp. 1–58.
7. Shakhnovich, E. I. & Gutin, A. M. (1990) *Nature (London)* **346,** 773–775.
8. Chan, H. S. & Dill, K. A. (1990) *J. Chem. Phys.* **92,** 3118–3135.
9. Chotia, C., Levitt, M. & Richardson, D. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4130–4134.
10. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1092.
11. Karplus, M., Caflisch, A., Šali, A. & Shakhnovich, E. (1995) in *Modeling of Biomolecular Structures and Mechanisms*, eds. Pullman, A., Jortner, J. & Pullman, B. (Kluwer, Boston), pp. 69–84.
12. Shakhnovich, E. I. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7195–7199.
13. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72,** 3907–3910.
14. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *J. Chem. Phys.* **101,** 6052–6062.
15. Karplus, M. & Šali, A. (1995) *Curr. Opin. Struct. Biol.* **5,** 58–73.
16. Ueda, Y., Taketomi, H. & Gō, N. (1978) *Biopolymers* **17,** 1531–1548.
17. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990) *Science* **247,** 1306–1310.
18. Desjarlais, J. R. & Handel, T. M. (1995) *Protein Sci.* **4,** 2006–2018.
19. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *Biochemistry* **33,** 10026–10036.
20. Govindarajan, S. & Goldstein, R. A. (1995) *Biopolymers* **36,** 43–51.
21. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995) *J. Mol. Biol.* **252,** 460–471.
22. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995) *Protein Sci.* **4,** 561–602.
23. Zwanzig, R., Szabo, A. & Bagchi, B. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 20–22.
24. Karplus, M. & Shakhnovich, E. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), pp. 127–193.
25. Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 1942–1946.
26. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1995) *Biochemistry* **34,** 3066–3076.
27. Brocchieri, L. & Karlin, S. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12136–12140.
28. Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1992) *Biochemistry* **31,** 10226–10238.
29. Avbelj, F. & Moult, J. (1995) *Proteins* **23,** 129–141.
30. Dyson, H. J., Merutka, G., Waltho, J. P., Lerner, R. A. & Wright, P. E. (1992) *J. Mol. Biol.* **226,** 795–817.
31. Radford, S. E., Dobson, C. M. & Evans, P. A. (1992) *Nature (London)* **358,** 302–307.
32. Hooke, S. D., Radford, S. E. & Dobson, C. M. (1994) *Biochemistry* **33,** 5867–5876.
33. Elöve, G. A., Chaffotte, A. F., Roder, H. & Goldberg, M. E. (1992) *Biochemistry* **31,** 6876–6883.
34. Bai, Y., Sosnick, T. R., Mayne, L. & Englander, S. W. (1995) *Science* **269,** 192–197.
35. Otzen, D. E., Itzhaki, L. S., ElMasry, N. F., Jackson, S. E. & Fersht, A. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 10422–10425.
36. Fersht, A. R. (1993) *FEBS Lett.* **325,** 5–16.
37. Creighton, T. E. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), pp. 301–352.
38. Weissman, J. S. & Kim, P. S. (1995) *Nat. Struct. Biol.* **2,** 1123–1130.
39. Sippl, M. J. (1995) *Curr. Opin. Struct. Biol.* **5,** 229–235.
40. Lemer, C. M.-R., Rooman, M. J. & Wodak, S. J. (1995) *Proteins* **23,** 337–355.
41. Karplus, M. & Weaver, D. L. (1994) *Protein Sci.* **3,** 650–668.
42. Sosnick, T. R., Mayne, L. & Englander, S. W. (1996) *Proteins Struct. Funct. Genet.* **24,** 413–426.