Specific Nucleus as the Transition State for Protein Folding: Evidence from the Lattice Model[†]

V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich*

Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge Massachusetts 02138

Received February 1, 1994; Revised Manuscript Received June 10, 1994[®]

ABSTRACT: We have studied the folding mechanism of lattice model 36-mer proteins. Using a simulated annealing procedure in sequence space, we have designed sequences to have sufficiently low energy in a given target conformation, which plays the role of the native structure in our study. The sequence design algorithm generated sequences for which the native structures is a pronounced global energy minimum. Then, designed sequences were subjected to lattice Monte Carlo simulations of folding. In each run, starting from a random coil conformation, the chain reached its native structure, which is indicative that the model proteins solve the Levinthal paradox. The folding mechanism involved nucleation growth. Formation of a specific nucleus, which is a particular pattern of contacts, is shown to be a necessary and sufficient condition for subsequent rapid folding to the native state. The nucleus represents a transition state of folding to the molten globule conformation. The search for the nucleus is a rate-limiting step of folding and corresponds to overcoming the major free energy barrier. We also observed a folding pathway that is the approach to the native state after nucleus formation; this stage takes about 1% of the simulation time. The nucleus is a spatially localized substructure of the native state having 8 out of 40 native contacts. However, monomers belonging to the nucleus are scattered along the sequence, so that several nucleus contacts are long-range while other are short-range. A folding nucleus was also found in a longer chain 80-mer, where it also constituted 20% of the native structure. The possible mechanism of folding of designed proteins, as well as the experimental implications of this study is discussed.

Theoretical studies of the thermodynamics and dynamics of protein folding have been reviewed recently in Karplus and Shakhnovich (1992). The authors pointed out that different approaches should be taken to study different parts of configurational space. The neighborhood of the native state and the dynamics of thermal fluctuations around this state can be studied in detail using an all-atom representation of a protein and applying molecular dynamics to simulate the system. The simplistic point of view would be to extend these calculations further to explore more of the configurational space and to also address the folding problem. However, this is impossible due to the obvious time limitations of such calculations. This implies that simplified models should be used to study folding. These models should be adequate to the problem, but free of details that are relevant on time and length scales much smaller than the ones at which interesting folding events occur. The adequacy of a model for the folding problem requires that model proteins possess a unique structure that is thermodynamically stable at physiological temperatures. The model should have the Levinthal paradox, i.e., an astronomically large number of conformations that cannot be scanned exhaustively in a folding simulation.

The idea of "preaveraging" irrelevant fast degrees of freedom leads to low-resolution models such as "beads on a string" (Lifshitz et al., 1978) or closely related lattice models (Ueda et al., 1978; Shakhnovich & Gutin, 1990a; Covell & Jernigan, 1990; Lau & Dill, 1989; Skolnick & Kolinski, 1990a,b). In such models, a group of atoms of a protein is represented by one effective monomer; one could visualize this as a C_α representation of protein folds. These models capture important aspects of the protein folding problem: an astronomically large number of conformations, the polymeric structure of the chain, and the chain heterogeneity [monomers (although represented by structureless "beads") may be of different types manifested by interresidue interactions of different strengths and signs]. The identity of a model protein is determined by the sequence of monomers. How can one study the folding of such model proteins? The key requirement is that simulations be unbiased to the native state and converge repetitively to one conformation independent of initial conditions—just as real proteins do.

The straightforward and desirable approach to folding proteins even within simplified models is to use the natural amino acid sequences of some moderately sized proteins and simulate their folding by Monte Carlo or molecular dynamics (Wilson & Doniach, 1989; Skolnick & Kolinski, 1990b). However, the major problem with this approach is that protein sequences may have been evolutionarily designed to satisfy folding requirements with a certain "exact" force field. Simulations necessarily use some approximate force field for which the native structure may be neither a global nor a pronounced kinetically accessible local minimum. When the force field is not completely adequate, the natural sequence is effectively random. Therefore, in order to explore this avenue, knowledge of the precise force field is necessary. The attempts to overcome this difficulty were based on the introduction of certain biases [e.g., making only the native contacts favorable (Ueda et al., 1978) or forcing the chain to acquire native secondary structure (Skolnick & Kolinski, 1990b)]. However, model Hamiltonians where such biases are introduced are somewhat unphysical.

A possible approach to unbiased simulations is to study short chains for which some subset of conformations can be enumerated. Then a nonspecific parameter, such as the average attraction between monomers, can be chosen in such a way that the global minimum would belong to this enumerated subset and therefore is known. Folding simula-

[†] E.I.S. was supported by the Packard Foundation.

[®] Abstract published in *Advance ACS Abstracts,* July 15, 1994.

0006-2960/94/0433-10026$04.50/0

tions would reveal whether this conformation of the global minimum is accessible or not.

This approach was taken by Shakhnovich et al., (1991) to study 27-mer chains on a simple cubic lattice and later by Miller et al. (1992) and Camacho and Thirumalai (1993) for 2-dimensional lattices. The folding of model proteins in these works was studied by lattice Monte Carlo simulations (Verdier, 1973; Hilhorst & Deutch, 1975). The important result obtained by Shakhnovich et al. (1991) is that folding and nonfolding sequences exist. The detailed analysis based on folding simulations for 200 random sequences (Sali et al., 1994) showed that the difference between folding and nonfolding sequences is that folding sequences had as their native structure a pronounced global energy minimum (Sali et al., 1994). Unfortunately, such an approach cannot be extended beyond 27-mer chains in three dimensions because the computational complexity of enumeration grows dramatically as chain length increases. However, it is clear that one should not necessarily enumerate all conformations; what is really very important is to know the global minimum conformation and its relation (in energy scale) to the multitude of remaining conformations.

This leads to the idea of combining design and folding to study the folding of longer protein size chains. The idea is simple: to design a sequence that will deliver sufficiently low energy to a given structure, so that one can be certain that this "target" structure represents a pronounced global minimum for this sequence. The specific choice of force field is not essential at this stage, provided that the design of a sequence satisfying the conditions mentioned above is possible with this force field. This sequence then can be subjected to a folding simulation with the same force field that was used at the design stage. At this point, one can hope that the simulation will converge to the target conformation for which the sequence was designed. The key idea here is to use the same force field for the folding simulation and for sequence design. This allows us to address the fundamental questions of protein folding separately from the practically very important but difficult question of which force fields are the most appropriate to study real proteins.

A step in this direction was made in a recent work by O'Toole and Panagiotoupoulos (1992) in which symmetric native structures and a simplified 2-letter, HP (hydrophobic, polar), representation of protein sequences were used. The design was based on the requirement to place more hydrophobic groups inside and hydrophilic groups outside. However, this attempt was not successful for longer chains since the designed sequences did not fold to their target structures. This is likely due to the deficiency of the 3-dimensional, 2-letter HP model, which does not have a stable unique conformation of the global energy minimum (Shakhnovich, 1994).

The idea of combining design and folding was realized successfully recently when an effective sequence design algorithm based on a Monte Carlo (MC) optimization procedure in sequence space became available (Shakhnovich & Gutin, 1993a,b). This made it possible to use a more realistic sequence representation of monomers of 20 types and allowed lattice model folding of proteins of different lengths (36–100) (Shakhnovich, 1994). This approach provides a unique opportunity to study the mechanism by which model proteins solve their folding problems, which is by no means simpler than that of real proteins. Indeed, the shortest of the model proteins we worked with is a 36-mer, which has $4.68^{35} \sim 10^{25}$ conformations (Sykes, 1963), too large a number to be scanned exhaustively. (For 100-mers, which also fold in our simulations, this number is 10^{75}!) Since we eventually can trace any intermediate conformation in the simulation, a very detailed study of the mechanism of folding can be done for the model proteins.

The statistical mechanics of proteins with designed sequences was discussed by Shakhnovich and Gutin (1993a), who showed that the sequences undergo a first-order folding transition to the native state. [For the qualitative explanation of the nature of first-order transitions in biomolecules, see Karplus and Shakhnovich (1992).] The phenomenological model of Bryngelson and Wolynes (1987), where the idea of design was encapsulated in the "principle of minimal frustration", also implied that transition to the native state may have first-order character. But the mechanism of first-order transitions is known to involve nucleation and growth (Lifshitz & Pitaevskii, 1981). Therefore, it is natural to expect the nucleation growth mechanism of protein folding.

The idea of a nucleation growth mechanism in protein folding was suggested by Levinthal in a largely unavailable publication (Levinthal, 1969) and was pursued in the subsequent work of Tsong et al. (1972) on the basis of kinetic analysis of experimental data and by Wetlaufer (1973) on the basis of observation of existing protein structures. The nucleation mechanism was also discussed in a recent work (Mault & Unger, 1992). In these works, the nucleation growth mechanism was based on phenomenological models, and detailed microscopic study to support or reject this hypothesis was missing. In this study, we suggest a detailed microscopic analysis on the basis of the lattice model of protein folding.

METHODS

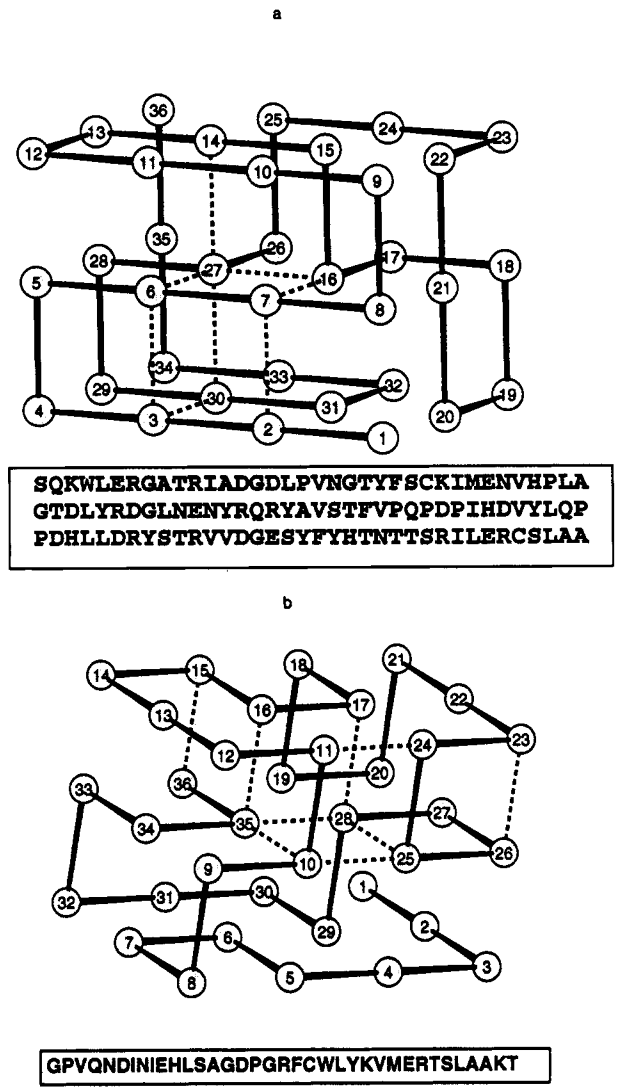
We use a 36-mer chain on a cubic lattice as a basic model (some results for longer chains will be sketched in the Discussion). We tried two different arbitrarily chosen compact native structures in order to determine which conclusions depend on the structural features of the native state and which are independent of it (Figure 1). The next step was to design a sequence that fits the native structure with a low energy. To this end, we used a MC optimization algorithm in sequence space, documented in detail by Shakhnovich and Gutin (1993a,b).

The energy function that we used throughout this work is taken in the nearest-neighbor approximation (Miyazawa & Jernigan, 1985):

$$E_0(\{\sigma_i\}) = \frac{1}{2}\sum_{i,j}^{N} U(\sigma_i\sigma_j)\Delta(r_i^{\text{native}} - r_j^{\text{native}}) \qquad (1)$$

where $N = 36$ is the total number of monomers and Δ defines the contact potential between them: $\Delta(r) = 1$ if $r_{\text{low}} < r < r_{\text{high}}$ and $\Delta(r) = 0$ otherwise. Our model protein is positioned on a simple cubic lattice with bond length of 3.8 Å. The target native conformation is set through the coordinates of its monomers $\{r^{\text{native}}\}$. Any two monomers that are 3.8 Å apart (so that, say, $r_{\text{low}} = 3.7$ Å and $r_{\text{high}} = 3.9$ Å) are considered to be in contact. For the set of potentials $U(\sigma_i\sigma_j)$, we used parameters determined by Miyazawa and Jernigan (1985) (MJ) from the statistical distribution of contacts in native proteins. The sequence design algorithm was run at low selective temperature [see Shakhnovich and Gutin (1993a)], $T_{\text{sel}} = 0.2$, to provide sequences that fit the native structure with sufficiently low energy.

The MC procedure in sequence space requires the initial setting of amino acid composition. We tried several choices. First we designed proteins with an "average" amino acid



SQKWLERGATRIADGDLPVNGTYFSCKIMENVHPLA
GTDLYRDGLNENYRQRYAVSTFVPQPDPIHDVYLQP
PDHLLDRYSTRVVDGESYFYHTNTTSRILERCSLAA

GPVQNDINIEHLSAGDPGRFCWLYKVMERTSLAAKT

FIGURE 1: Target conformations used in this study. Sequences are shown that fit the corresponding conformations with sufficiently low energy to make sure that these conformations are global energy minima for the designed sequences. We worked with three sequences designed for structure a and one sequence designed for structure b. Most calculations were done with structure a; however, for the sake of control the nucleus was also determined for structure b. Dashed lines denote contacts belonging to the nucleus.

composition taken from Table 1.1 of Creighton (1992). Another choice was to take a composition like that of the small 36-residue protein pancreatic bird polypeptide (1 ppt).

Since we are using MJ parameters that were obtained from protein statistics, we have only relative energy and do not know the absolute energy scale for this set of parameters (Finkelstein et al., 1993). So we use the energy unit at which DU = 1, where $\mathrm{DU} = ((\langle U^2\rangle - \langle U\rangle^2))^{1/2}$ is the standard deviation of the energy of different interactions; this is the measure of their heterogeneity. $\langle\ \rangle$ denotes averaging over all possible pairwise interactions in the given sequence:

$$\langle U^{\mathrm{p}}\rangle = \frac{2}{N(N-1)}\sum_{i\leq j}^{N} U^{\mathrm{p}}(\sigma_i\sigma_j) \qquad (2)$$

The design procedure generated a number of sequences; we intensively studied the ones shown in Figure 1.

The lattice Monte Carlo simulations of the folding of designed sequences are done with a standard algorithm well documented in earlier works (Verdier, 1973; Hilhorst & Deutch, 1975; Sali et al., 1994). The standard move set was taken to include corner flips and crankshaft motions (Hilhorst & Deutch, 1975). The Metropolis criterion with the energy function (eq 1) was used (Metropolis et al., 1953) to accept or reject moves.

To measure the structural similarity between a current conformation and the native state, we used the similarity parameter Q (Shakhnovich & Gutin, 1989a,b, 1990a), which is the normalized number of native contacts in a conformation:

$$Q = \frac{N_{\mathrm{native}}}{N_{\mathrm{total}}}$$

where N_{total} is the number of contacts in the compact conformation; $N_{\mathrm{total}} = 40$ for the 36-mer. It follows from this definition that $Q = 1$ in the native state.

Simulations started from the random coil (see an example in Figure 2) and ended when the native target structure was reached (Figure 3). The mean first passage time for reaching the native state was $\sim 10^6$ Monte Carlo steps at $T = 0.90$, at which all simulations reported in this work were performed. The native conformation (shown in Figure 1a,b for corresponding sequences) had the lowest energy among all conformations found in the simulations. To test this, a long simulation of 10^9 Monte Carlo steps was run to make sure that no other structures with energy equal to or lower than the energy of the native structure were encountered. This was indeed the case, which made us sufficiently confident that the native is the global minimum of energy.

SEARCH FOR THE NUCLEUS

Exploring implications of the first-order transition kinetics of folding we expect that the chain overcomes the main free energy barrier via a nucleation growth mechanism. There are two slightly different definitions of nuclei in the kinetics of the first-order transitions (Lifshitz & Pitaevskii, 1981). The critical nuclei correspond to transition states (free energy barriers). There is a probability of roughly 1/2 that the new phase will grow further after the critical nucleus is formed and a probability of 1/2 that it will dissolve. One can also define a postcritical nucleus, i.e., the minimal sized fragment of the new phase that inevitably grows further to the new phase. Certainly there is no great difference between the two ways of defining the nucleus because the postcritical nucleus simply should have energy a few $k_{\mathrm{B}}T$ lower than the critical one, the barrier state, in order to make the subsequent growth unidirectional and irreversible. In our study, we will be interested in postcritical nuclei, i.e., ones that subsequently grow into the folded state.

The main difficulty in finding a nucleus comes from the fact that they are very short-lived before they grow further into the native or near-native conformation. By no means should they be confused with intermediates that are long-lived and detectable because they are sufficiently deep local minima. We define a nucleus as a set of contacts that satisfies the following two conditions: (i) Formation of a nucleus is a sufficient condition for folding; i.e., after a set of contacts that constitutes the nucleus is formed, the subsequent folding is guaranteed and is very fast (in our search for a nucleus we required that folding should take place in less than 50 000 MC steps after the nucleus is formed). We are therefore looking for postcritical nuclei. (ii) Formation of a nucleus is a necessary condition for folding; i.e., the pattern of contacts corresponding to the nucleus is *always* present in "prefolding conformations" when the number of native contacts is relatively small, but subsequent folding is very fast.

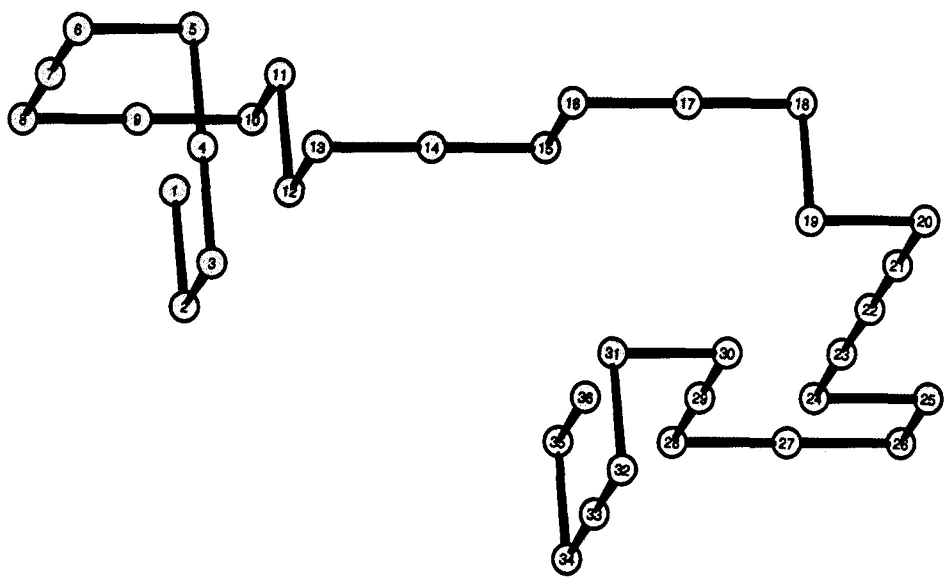


FIGURE 2: Example of a starting random coil conformation.

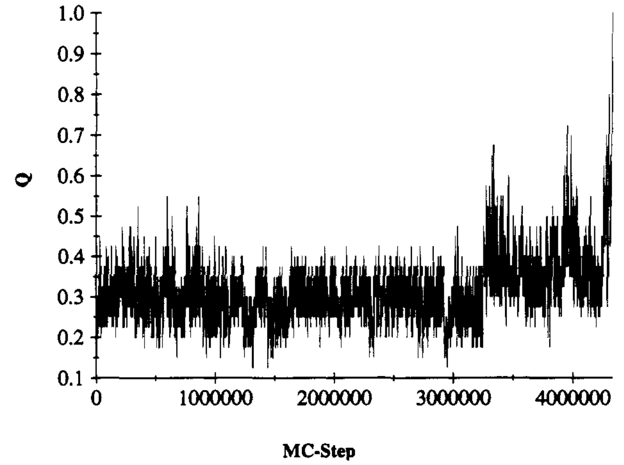


FIGURE 3: Example of a folding trajectory starting in the random coil conformation and ending in the native state. Such types of trajectories were used throughout this study.

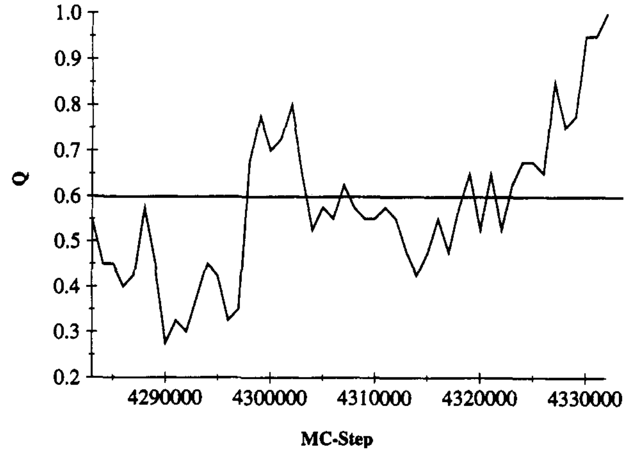


FIGURE 4: Part of the folding trajectory from Figure 3 used to search for the nucleus. The horizontal line illustrates the criterion $Q < 0.6$ for the choice of conformations relevant to the search for the nucleus.

The last condition requires some explanation. It is trivial that in the vicinity of the native state where $Q \approx 1$, some contacts will consistently appear just before the native state is reached. What we are interested in is the *minimal* set of contacts that *must* be formed before folding proceeds to the native state. To this end, we should analyze conformations that are not too close to the native state. As inspection of Figure 3 suggests, in the largest part of the trajectory the chain is fluctuating in conformations with Q not exceeding 0.6. This means that we should search for nuclei by analyzing sets of contacts that are present in conformations belonging to steep parts of the trajectory (Figure 4) but that are structurally different from the native state. To this end, we analyzed all conformations with $Q < 0.6$ (see Figure 4) that are separated by less than 50 000 Monte Carlo steps from the final step of the simulation when the native state was reached. The data were collected over 10 runs, each starting from a random coil and ending in the native conformation. Our analysis was aimed at revealing the set of contacts common to all 10 runs.

We discovered that rapid folding always takes place after the formation of a distinct set of eight contacts (shown by dashed lines in Figure 1a) for the first target structure and nine contacts for the second target structure (Figure 1b). We can see that contacts forming the nucleus are located in the native structure in the vicinity of each other, not in random positions. These contacts form a spatially localized substructure, which serves as a nucleus for folding. The formation of this set is indeed both necessary and sufficient for fast folding. It is the necessary condition because this set of contacts is formed for the first time only several thousand Monte Carlo steps before the native state is reached and in conformations for which the number of native contacts is relatively small (less than 25 out of 40). It was also a sufficient condition because after the nucleus had been formed the native state was always reached in less than 50 000 Monte Carlo steps, or about 1% of the total Monte Carlo time of folding from a random conformation.

Another important finding was that the position of the nucleus was nonspecific to the sequence chosen: for all three sequences shown in Figure 1a, the position of the nucleus was the same. We analyzed folding trajectories for 30 more sequences designed to have the native structure, as shown in Figure 1a, and found that in all these trajectories formation of the nucleus shown in Figure 1a preceded subsequent fast folding to the native state. To avoid confusion here, we should stress that although these sequences are nonhomologous, they are not independent either: they were all designed to have the structure shown in Figure 1a as the global minimum conformation.

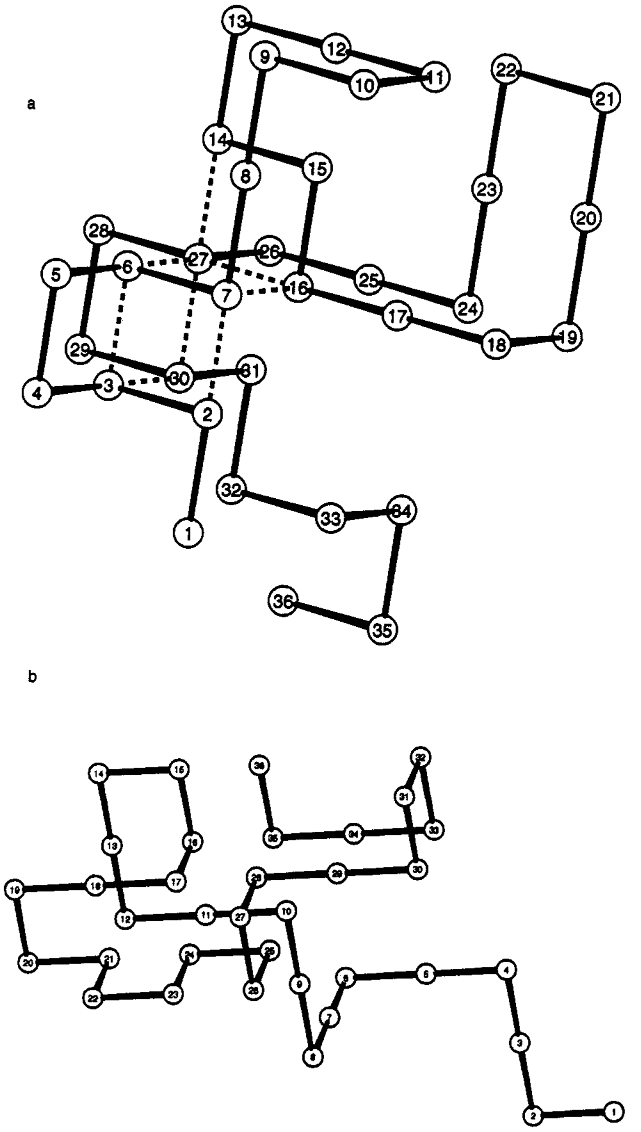


FIGURE 5: (a) Example of a starting conformation containing nucleus contacts. Otherwise the conformation was random. (b) Control: Starting conformation containing the same number of native-like contacts as in a, but without nucleus contacts.

EXPLORING THE NUCLEATION MECHANISM

As the first test of the proposed nucleation mechanism, we studied folding trajectories that started from a conformation with a preformed nucleus; otherwise this conformation was completely random and noncompact (see Figure 5a). It contained about ten native contacts (eight belonging to the nucleus and two randomly formed). Therefore, $Q \approx 0.25$ in a starting conformation. When simulations were started from conformations with the preformed nucleus, as shown in Figure 5a, the native state was reached quickly (on average in less than in 30 000 MC steps, and in many runs in less than in 1000 MC steps). The time course of a typical simulation, which started from a conformation with a preformed nucleus is shown in Figure 6.

However, the question may arise whether fast folding from a conformation with a preformed nucleus is due to formation of the nucleus or whether any eight native contacts in the starting conformations provide such fast folding. In order to address this issue, we ran a control experiment starting from several conformations that contained at least eight native contacts but that were different from the nucleus ones (see an example of such starting conformation in Figure 5b). In all of these control experiments, the folding trajectories were practically indistinguishable from the ones that started from completely randomized conformations (Figure 3). The folding time distribution was unaffected by the choice of initial conformation in this case and yielded the same mean first passage folding time as before of close to 1 million MC steps. This can be rationalized if we look at any arbitrary trajectory that starts from a random coil (Figure 3). In fact, 8–10 native-like contacts (i.e., conformations with $Q \approx 0.2$–0.25) are formed at the very beginning of the simulation (in less than 100 000 MC steps). However, this does not lead to rapid folding: a few million more steps are required to reach the native structure. Only formation of the *specific* subset of contacts, the nucleus, results in rapid folding.

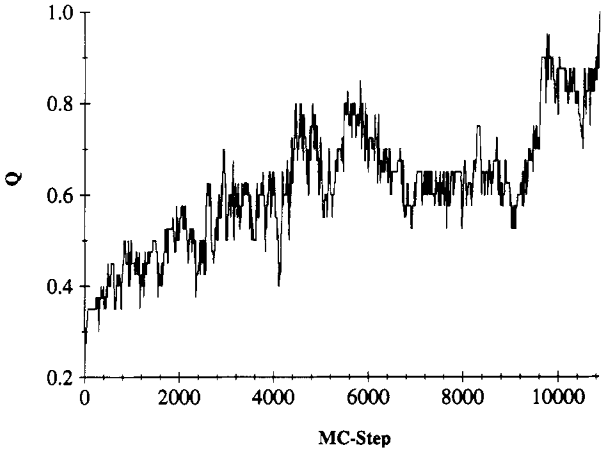


FIGURE 6: Typical folding trajectory for runs that start from a conformation with a preformed nucleus, as shown in Figure 5a.

As was mentioned before, formation of the postcritical nucleus corresponds to the transition over the main free energy barrier. This implies that there must be a significant difference in folding mechanism when the simulations start from completely randomized conformations and when they start from a conformation with a preformed nucleus, as shown in Figure 5a. In the first case, one should expect that the rate-limiting stage is overcoming the main barrier or formation of the nucleus, while in the latter case the motion to the native state would be downhill in free energy space, representing an effective pathway or funnel (Leopold et al., 1992).

To test this, we compared the statistical characteristics of the folding process in both cases. In the case where folding started from a random conformation, we evaluated at each trajectory, after each 1000 MC steps, the number of all current contacts (y) as well as the number of the native contacts (x). The frequency with which specific pairs (x,y) were found in 10 folding trajectories was evaluated to calculate the probability $P(x,y)$ of finding a conformation with y contacts, x of which are the native ones. These results are illustrated in Figure 7a. Both x and y can take values from 0 to 40, and $41^2 = 1681$ dots correspond to 1681 possible pairs of x and y. The higher $P(x,y)$, the lighter the corresponding dot on Figure 7.

We should note here that our experiments were aimed at the estimation of the mean first passage time, and therefore simulations ended when the native conformation had been reached. This explains the apparent low population of the native state in Figure 7. In fact, the native state was rather stable at that temperature, having $\langle Q \rangle \approx 0.8$ where $\langle\ \rangle$ denotes thermal averaging over long (equilibrium) trajectories.

One can see that conformations with approximately 25 total and 15 native contacts are most frequent. This is certainly

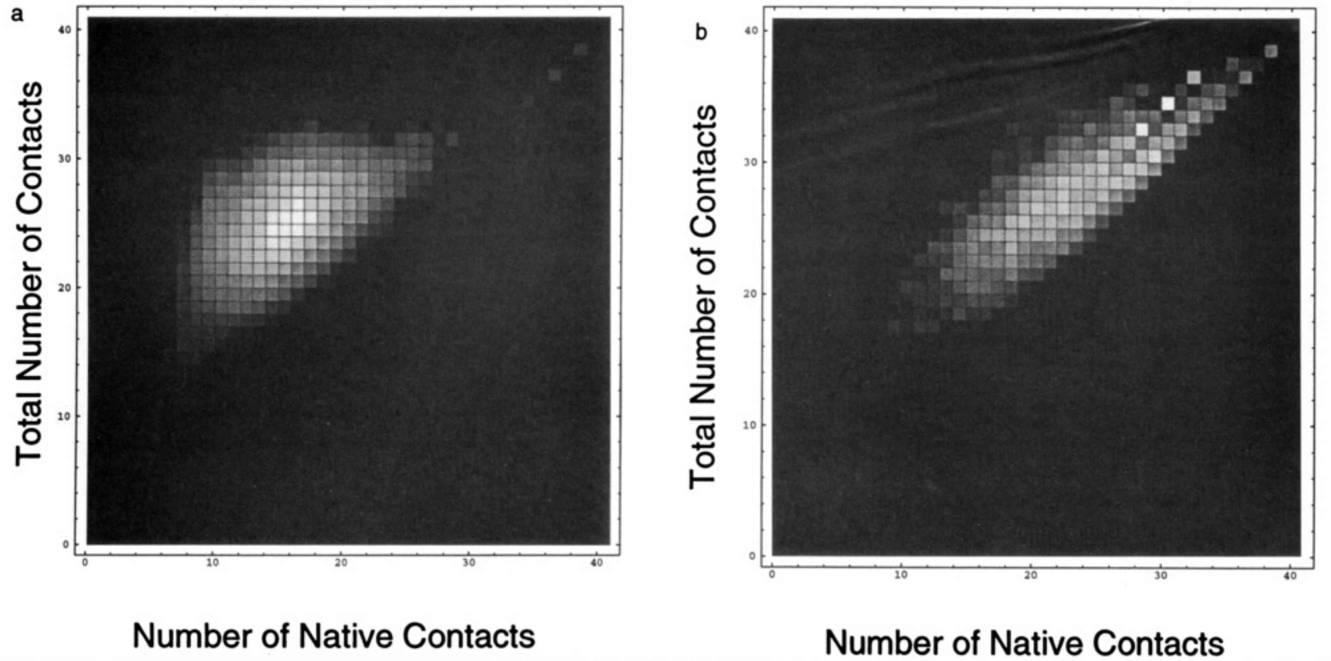


FIGURE 7: Density plots illustrating the frequencies with which conformations having a specified number of native contacts (abscissa) and total number of contacts (ordinate) are found in simulations. The brighter the dot with coordinates (x,y), the more frequently conformations with x total and y native contacts were found. (a) Simulations starting from completely random conformations; (b) simulations starting from conformations with a preformed nucleus, as shown in Figure 5a.

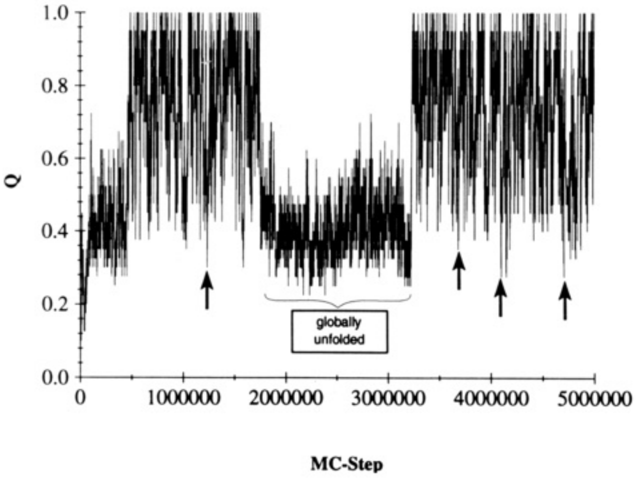


FIGURE 8: Part of a longer trajectory containing local unfolding events shown by arrows and global unfolding shown by the bracket. Local unfolding is as deep as the global one; however, locally unfolded conformations usually refold in less than 20 000 MC steps.

a prebarrier minimum of free energy or a folding intermediate. Conformations with more than 30 native contacts are rare. This means that the chain spends most of its folding time fluctuating around the intermediate state until it reaches a conformation(s) corresponding to the free energy barrier, after which folding is fast. The computer experiments described earlier show that this is the set of conformations containing the nucleus.

The same calculations were performed when folding started from conformations that contained a preformed nucleus, as shown in Figure 5a. The only difference was that the numbers of native and all contacts, x and y, were evaluated at each tenth Monte Carlo step because the folding time was substantially smaller. The results are illustrated in Figure 7b. There is a clear difference between the plots shown in Figure 7a,b. In the case of folding from a preformed nucleus, the number of native contacts is very strongly correlated with the number of all contacts, as the light area on Figure 7b is stretched along the main diagonal $x = y$. It is also important to note that there is no maximum of $P(x,y)$ on Figure 7b, and we can see that $P(x,y)$ is approximately constant in the area close to the diagonal $x = y$ and vanishes everywhere else, which implies that in this case the chain is not wandering randomly through conformational space but folds quickly, increasing the number of native contacts at an approximately constant rate (a clear indication of the propagation mechanism). Of course the polypeptide chain still has a tremendous number of conformations, but the constant value of $P(x,y)$ suggests that a directed assembly takes place after the nucleus is formed. Thus, the addition of any native contact decreases free energy, and this driving force directs the process. No significant free energy barriers are found in this part of the configurational space.

We studied the role of the nucleus in the initiation of folding in our model. However, for conformations with a nucleus, proximity to the transition state may also play an important role in unfolding. To this end, we studied longer trajectories, during which several folding–unfolding events occurred (Figure 8). Inspection of these trajectories reveals two possible scenarios of transient unfolding. The first type of behavior corresponds to significant unfolding (up to $Q \approx 0.2$), but after 10000–20000 MC steps the chain refolded back. Such unfolded conformations after which the chain refolds quickly will be called locally unfolded. However, sometimes the same degree of unfolding to $Q \approx 0.2$ led to more dramatic consequences: a few million MC steps were required for the chain to refold (see Figure 8). Conformations that required such a long time to refold will be called globally unfolded.

The question then is what is the difference between globally and locally unfolded conformations? We studied 10 different long (up to 100 million MC steps) folding trajectories (part of one is shown in Figure 8) and examined all locally unfolded conformations with less than 16 native contacts. We found that all of these conformations contained the intact nucleus, while globally unfolded conformations missed contacts from the nucleus. An implication of this observation is that although fluctuations in the folded state are significant, some contacts

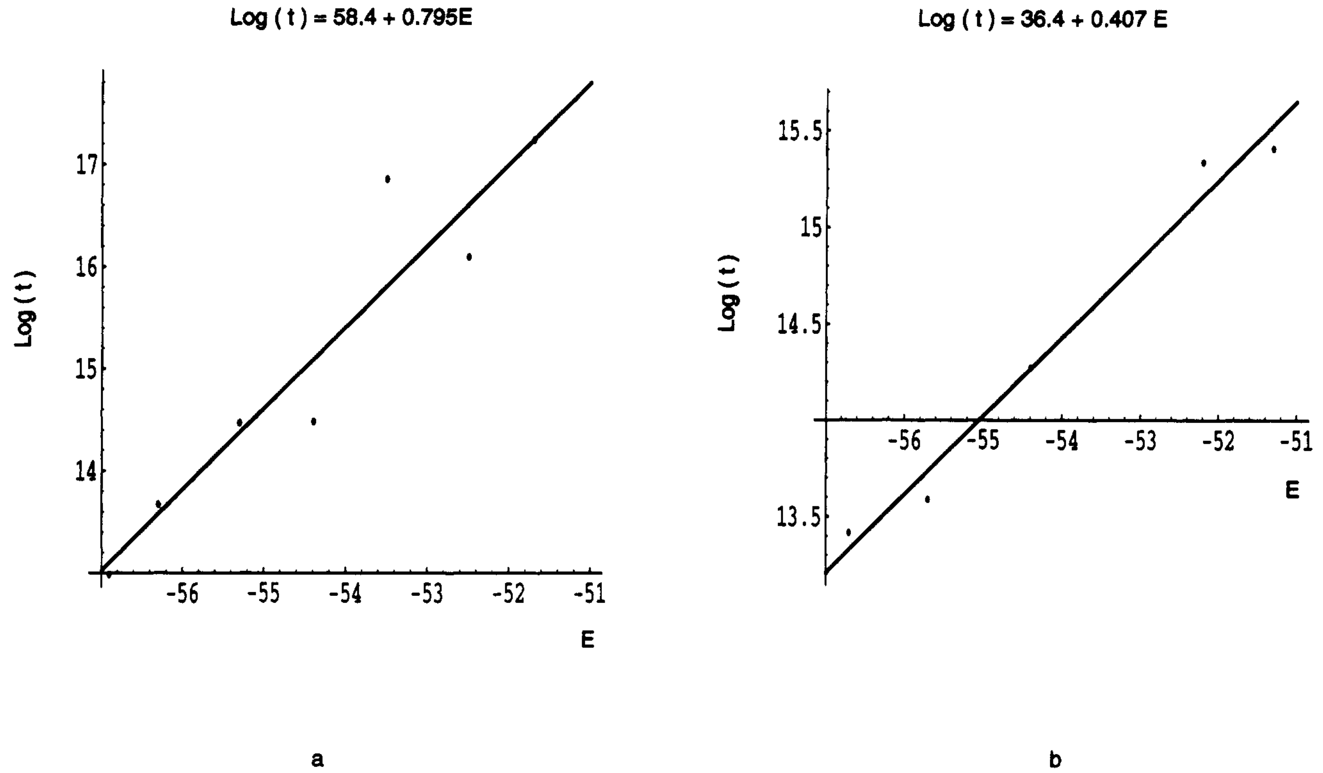


FIGURE 9: (a) Dependence of the log of MFPT of folding on the energy of the native state for a number of sequences having different energy of nucleus contacts and the same total energy of the other contacts. (b) Same dependence but for the set of sequences having the same energy of nucleus but different total energy of remaining contacts.

are more stable than others—a clear indication of the heterogeneity of the folded conformation in our model, which we relate to the molten globule state in proteins (see the Discussion section). This result shows also that there is no other nucleation site in our chains. If there were, we would see either of the nucleation sites preserved in locally unfolded conformations, but we definitely see (repetitively) only one subset of contacts common to all locally unfolded conformations. We note, however, that this conclusion is drawn for 36-mer chains, and it is certainly possible that longer chains may have multiple nucleation sites. A very interesting question, then, is at what size of the chain (if any) the nucleation regime changes from one nucleus to multiple nuclei.

The next issue we addressed was the dependence of the folding time on the interaction energy of the contacts constituting the nucleation site. To this end, using the same design procedure (Shakhnovich & Gutin, 1993a,b), we selected a set of sequences having a different total energy for the 8 nucleus contacts but a similar total energy for the remaining 32 native contacts. The objective was to study how the stability of the nucleus affects the rate of folding. The result is presented in Figure 9a, where dependence of the logarithm of the folding time on the total energy of the native conformation, normalized by kT, is shown. We would like to emphasize that although we plot the mean folding time vs the *total* energy of the native conformation, the sequences corresponding to the different data points in Figure 9a differ by the energy of nucleus contacts only, having similar energy for the remaining contacts in the native conformation. The dependence presented in Figure 9a is close to linear with a slope of 0.8. This should be contrasted with the results of a control experiment in which sequences were chosen to have similar energy for nucleus contacts and differ in energy for the remaining contacts (Figure 9b). In this case, the dependence of log(time) on the energy of the structure is also close to linear, but the slope is half as great (0.4). This indicates that stabilization of the nucleus is more important for rapid folding than the stabilization of other contacts, although the latter may indirectly stabilize the nucleus, decreasing the entropic cost of its formation. This gives rise to the acceleration of folding in that case.

DISCUSSION

In this section, we will discuss two aspects of the present study. First, we discuss the lattice model results and their implications. In the second part of the Discussion, we will discuss the applicability of simplified lattice models to the study of the folding of real proteins: features that lattice models catch and features that they miss.

In this paper, we have provided a body of evidence that the folding mechanism of lattice proteins involves the formation of a *specific* nucleus as a *transition state*, with its subsequent growth. This is not at all unexpected because nucleation growth is a standard mechanism of cooperative (first-order) transitions; for instance, the vapor–liquid transition is well known to involve a nucleation growth stage (Lifshitz & Pitaevskii, 1981). There is, however, an essential difference between the nucleation growth mechanism in simple liquids and that in model proteins. In liquids a nucleus is nonspecific and is fully characterized by its size. In model proteins the nucleus is *specific*, which means that a particular set of contacts, constituting a *transition state*, should be formed to cause subsequent fast folding to the native state.

The folding process in each molecule involves two stages, which we can characterize as stochastic and deterministic. The stochastic stage is rate-limiting (the stage at which the nucleus is formed via random search). Of course this does not imply that the protein should "wait" for a multiparticle collision to form the nucleus. Since the nucleus is a substructure of the native state, its contacts are attractive and therefore the partly formed nucleus does not disappear. The possibility of a stochastic search to form nuclei was pointed out by Wetlaufer (1973). The stochastic search for the nucleus takes place in the intermediate that is formed at the burst stage of the folding process (in less than 30 000 MC steps). This burst intermediate can be seen as a light area on Figure 7a as a partly compact state (having 20 out of 40 contacts, 10–12 of which are the native ones). This intermediate represents a multitude of rapidly interconverting conformations, corresponding to a prebarrier free energy minimum. Formation of a burst semicompact intermediate precedes the formation of a nucleus, which is formed later when native contacts in the intermediate include, for the first time, the nucleus ones.

The subsequent folding, after the nucleus is formed, is fast and practically unidirectional. This is not surprising because formation of a nucleus is equivalent to overcoming the main free energy barrier. We observe *a folding pathway* that is the postnucleus assembly of the protein associated with the directed motion downhill in free energy. Indeed, as inspection of Figure 7b suggests, the number of native contacts grows steadily with the increase of the total number of contacts, i.e., roughly speaking, in this regime every added contact is a native one. It can also be seen that the chain does not encounter, at this temperature, significant barriers as it progresses through the pathway; the motion in configurational space is rather diffusion-like. The evidence for this is the approximate constant density in the light region of the diagram of Figure 7b.

It was suggested in previous works, implicitly (Wetlaufer, 1973) or explicitly (Rooman et al., 1992a,b), that at least a considerable part of the nucleus should be formed by contacts between residues that are close to each other in sequences (local contacts). Our analysis is consistent with these assertions. Inspection of Figure 1 shows that the nucleus is formed by both long-range as well as short-range contacts, with some predominance of the long-range contacts. However, the relation between the numbers of short-range and long-range contacts in the nucleus may depend on the potential chosen since the local component of the potential may increase the number of local contacts in the nucleus. This question requires further study. We believe, however, that some long-range contacts must always be present in the nucleus since such contacts are most effective in decreasing entropy of the transition state and thus creating an "entrance" to the pathway.

The results reported in the present paper were obtained for the 36-mer model proteins. A very important question is whether these results are valid for longer sequences. Our approach allows for folding longer sequences (at least up to 100-mers) (Shakhnovich, 1994a,b). To test the conclusions of a,b this work, we studied a nucleation mechanism of folding for a 80-mer chain. Using the same procedures as described in this work earlier, we found the nucleus for the 80-mer to have 22 out of 105 contacts. This included 16 monomers. Although we observed a single nucleus for the 80-mer chains, we cannot exclude a multiple-nuclei mechanism for longer chains. These multiple nuclei could be associated with folding domains [observed recently in hen lysozyme (Miranker et al., 1991)], which may or may not develop into the structural domains of native proteins.

The folding of long chains (36–100 residues) was possible only because these sequences were designed to have the native state as a pronounced energy minimum, and a special design procedure was necessary to generate such sequences (Shakhnovich & Gutin, 1993a,b). Long random sequences were not able to fold (Shakhnovich, 1994a). A complementary approach to study folding was taken in the recent paper by Sali et al. (1994b), where *short random* sequences were taken to study the "minimal requirements" for "one-shot" selection of folding sequences from the pool of random sequences. Analysis of the folding of short quasirandom folding sequences also revealed an activation mechanism, but it differs from the one found in the present study. The transition state for short random sequences turned out to contain 80–95% of native contacts (compare with the nucleus that has 8 particular contacts out of 40). This difference may be due to a number of reasons. First of all, a random interaction energy model was studied by Sali et al. (1994b), while here we studied a more realistic sequence model. This difference may be important since in the former model energies of contacts are totally uncorrelated, while in the latter energies of different contacts are correlated. Indeed, the identity of a protein is characterized in the letter model by N "letters" (primary structure), which determines $\sim N^2$ interactions between every pair of monomers. This implies that these interaction energies cannot be independent. In the random interaction energy model, the identity of a protein is defined through setting all $\sim N^2$ interactions between any pair of contacts independently. Correlations may be important for nucleus formation, which is a contiguous subset of native, stable contacts. A second reason, which is more likely to explain the difference between the results of two models, is that present sequences were designed to enable the folding of long chains. It is likely that design in the sequence model generated a contiguous subset of strong contacts, which turned out to be a nucleus. It was pointed out by Sali et al. (1994b) that the model used there is likely to describe the folding of prebiological, short, and poorly optimized sequences. As longer proteins evolved, their folding may have required sequence design that developed a more effective nucleation growth mechanism. Indeed, the characteristic folding "time" of random 27-mers in Sali et al. (1994b) was 20–50 million steps, while in the present study 36-mers fold in 1–5 million steps and designed 80–100 mers fold in 5–10 million MC steps (Shakhnovich, 1994a,b).

The results reported in this paper were obtained using Monte Carlo simulation in the lattice model. Two questions are in order now: how representative is Monte Carlo for the kinetics of folding, and what is the relationship between lattice models and real proteins?

A comprehensive study of the role of lattice and move sets in the apparent dynamics of a polymer was performed by Skolnick and Kolinski (1990a, 1991), who showed that there is no significant dependence of observed dynamics on the choice of lattice (diamond or 210) or move set. Moreover, Rey and Skolnick (1991) compared the simulation results obtained by Monte Carlo on the simplest (diamond) lattice and by off-lattice Brownian dynamics. Their conclusion is that the main dynamic features observed are independent of the simulation technique chosen. It was shown also by Skolnick and Kolinski (1990) that the choice of local moves only, being most natural, provides the most realistic time scale picture, as judged in comparison with the master equation calculation.

Thus, in our view, the Monte Carlo approach (taking into account its computational effectiveness) may be plausible for depicting key features of kinetic processes associated with protein folding. However, it is unlikely that MC simulations can provide a description of all of the microscopic details of the process. Rather, general features, which are observed over thousands of steps, are of interest. This is the case in our study in which we focus on nucleus formation that takes place in 10^6 steps.

The most important question concerning the approach taken in this study concerns the relationship between lattice model proteins and real proteins. There is one obvious feature of real proteins that our model misses. This is the presence of side chains with their degrees of freedom and tight packing in the native state. Therefore, our model is aimed toward describing stages (if any) of the protein folding process that do not include the tight packing of side chains. It is widely believed now that packing of side chains occurs at the transition from molten globule (MG) to native (N) conformation. Experimental evidence has been accumulating (Williams et al., 1991; Hughson et al., 1990; Peng & Kim, 1994) that the molten globule, when at equilibrium, retains a significant part of the native-like backbone fold, in accord with theoretical predictions (Shakhnovich & Finkelstein, 1982, 1989; Finkelstein & Shakhnovich, 1989). It was suggested (Ptitsyn, 1973, 1987) that a "native-like" molten globule may be a universal intermediate on the protein folding pathway. Subsequent experimental findings (Ptitsyn et al., 1990; Matouschek et al., 1990, 1992; Jennings & Wright, 1993) strongly buttressed this point, providing evidence [especially in Jennins and Wright (1993)] that the transient long-lived intermediate is structurally close to the equilibrium "native-like" molten globule.

The folded state in our model should be related to the "native-like" molten globule. It is interesting to note that a chain in this state fluctuates around the native fold, but these fluctuations are inhomogeneous (nucleus contacts are fluctuating less than other contacts) (see Figure 8). This is in accord with experimental information about the molten globule (Hughson et al., 1990; Baum et al., 1989).

The nucleus transition state that we observe in this work is the transition state between a coil, or a structureless compact intermediate without unique structure (Elove et al., 1992; Radford et al., 1992), and the molten globule with elements of native-like fold. By no means should it be confused with the transition state between the native state (N) and the molten globule (MG), which is usually associated with the transition state for folding because the MG–N transition is the rate-limiting step for the whole process. This transition N–MG state is known, both from theory and experiment (Segawa & Sugihara, 1984; Shakhnovich & Finkelstein, 1989; Bycroft, 1990), to be close to the native state, differing from it by some small expansion [so small that protein core is mainly inaccessible to the solvent: see Segawa and Sugikhara (1984) and Matouschek et al. (1992)].

A significant simplification of the model is that it did not include explicitly secondary structure segments, which are stabilized by H-bonds and are able to move as a whole. This question is related to the secondary structure framework and related diffusion–collision hypotheses of folding (Kim & Baldwin, 1982; Karplus & Weaver, 1976). The physical mechanism assumed in these hypotheses is that native-like secondary structure is formed at early stages so that subsequent folding includes movements of segments as a whole, without their restructuring due to long-range interactions. This may give a kinetic advantage because the degrees of freedom associated with secondary structure become frozen, and the remaining search is feasible because it includes far fewer conformations. Therefore, in order to facilitate kinetics, secondary structure elements, after having been formed at the ultrafast stage of folding, should be so stable that their characteristic folding–unfolding interconversion time in the absence of long-range interactions is longer than the time of formation of long-range contacts [in the millisecond time range (Radford et al. 1992; Bycroft et al., 1990; Jennings & Wright, 1993)]. The only way to increase the interconversion time from basic nanoseconds to milliseconds, which is consistent with the second law, is to increase the stability of the helix. This requires ~10 kcal/m/helix of stabilization, which implies that the Boltzmann probability of such a stable isolated helix will be very close to 1. Recent studies of isolated fragments of myoglobin corresponding to helical segments in its native secondary structure did not lend evidence supporting the suggestion that isolated helixes are stable in the absence of long-range interactions (Waltho et al., 1993; Shin et al., 1993). Certainly, some fluctuating elements of native-like and nonnative secondary structure may form quickly. However, it is unclear (at least to us) how the formation of marginally stable fluctuating α-helixes and β-strands, with their degrees of freedom in equilibrium with all other degrees of freedom, can provide any *kinetic* advantage leading to the resolution of Levinthal's paradox.

Of course our calculations cannot rule out the framework-type mechanism because movements of helixes or β-strands as a whole are not included in the move set. However, what they show is that this mechanism, even if valid, is not the only, or necessary, way to solve the Levinthal paradox. Our calculations give an *example* that the protein folding problem, at a model level, can be solved without a framework-type mechanism.

The sequences we worked with in this model were designed to have the native conformation as a pronounced global *energy* minimum. The question is how can this optimization be related to real proteins. First of all, we note that a pronounced energy gap between the native state and the set of nonnative conformations is a *necessary* thermodynamic condition of the uniqueness of the native structure; this is independent of the model or the potential function chosen. The native structure must be thermodynamically stable at physiological temperature. This can be guaranteed only if the gap between the native structure and nonnative conformations is sufficiently large, i.e., many kT (Shakhnovich & Gutin, 1990). In other words, a large energy gap protects a unique structure from destruction by thermal fluctuations. However, our results go further and suggest that a pronounced energy gap is also a *sufficient* condition for sequences to fold rapidly to the native conformation.

These considerations do not contradict the fact that proteins are not highly stable. Experimental results (e.g., Privalov, 1979) suggest that the temperature of denaturation for most proteins is not too high, and therefore the difference in free energy between the native conformation and denatured states is moderate: 10–12 kcal/mol for a 100-residue protein at physiological temperature (Privalov, 1992). In order to give a correct interpretation of the thermodynamic data on protein stability, one should note that what is known to be small is the difference in *free energy* between the native and denatured states; this includes the entropic contribution. Energy differences between the native and denatured states are much more pronounced, as measured by the latent heat of denaturational transition and its cooperativity. The entropic factor is also essential for lattice proteins, making the unfolding temperatures not too high ($\approx$1.1 in our energy units) and the lattice proteins marginally stable, like real ones.

CONCLUSION: IMPLICATIONS FOR EXPERIMENT

In this study, we have presented a minimal theoretical model of protein folding. The model is free of internal inconsistencies and unphysical assumptions. Indeed, the simulations are not artificially biased toward the native state: all the chain "knows" when the simulation starts from a random coil conformation is the amino acid sequence. The Hamiltonian is physical: the interaction of, say, glycine with another glycine depends only on the spatial distance between the residues and does not depend on their positions in the chain or in the native conformation. Model proteins resolve the Levinthal paradox, exhibiting fast folding to the unique global minimum conformation without scanning the astronomically large number of possible conformations. We presented the possible mechanism.

Like any theoretical work, this one deals with a simplified representation of proteins, and the adequacy of the model for the system it studies is at issue. The only nontautological way to estimate the adequacy of a model is to formulate its predictions and compare them with experiment.

It should be noted here that the model studied in this paper represents a generic protein and is aimed toward the study of *universal* features of protein folding unrelated to the specific structural features of a protein molecule.

The theoretical analysis presented in this work has several implications directly related to experiment, as follows.

(1) The cooperative character of the coil–molten globule transition in natural (i.e., evolutionarily optimized) protein sequences contrasted with the non-cooperative character and absence of unique structure in the randomized sequence. This explains the difference in experimental results for proteins [bovine carbonic anhydrase and staphylococcal β-lactamase (Uversky et al., 1992) and staphylococcal nuclease (Gittis et al., 1993)], where "all-or-none" transitions were reported, and for the quasirandom sequence of the F2 fragment of tryptophan synthase (Chafotte et al., 1991), where the transition is non-cooperative.

(2) Theory demonstrated the heterogeneity of the folded state (in the context of our model, molten globule), asserting that some contacts (in-nucleus) are less subject to fluctuations than other contacts (off-nucleus) (see Figure 8 and the discussion there). Corresponding, the nucleus contact interconversion rate is much slower, as is manifested in higher HD protection factors. Such heterogeneity in protection factors in the molten globule was indeed observed in a number of proteins (Hughson et al., 1990; Jeng et al., 1990). The explanation is simple: conformations with the nucleus correspond to the top of the barrier, the transition state. Therefore, fluctuations that go up to the barrier are most rare as they require higher energy. This makes the nucleus the most protected region in a molten globule.

(3) Our calculations predict a direct correspondence between the residues that are most protected from HD exchange in the equilibrium molten globule and the ones involved in folding the nucleus, i.e., *the first* stable set of contacts to be formed in the course of the folding process. This assertion is in accord with the experimental results for myoglobin (Jennings & Wright, 1993) and cytochrome C (Roder et al., 1988). The observation about the implications of mutations in nuclei on the folding rate makes this correspondence directly experimentally verifiable.

Our design–folding approach provided a possible conceptual framework to solve the protein folding problem. Within this approach, one can also address questions pertinent to the folding pathway of a specific protein, e.g., how to determine the folding nucleus in a given protein. To this end, it is necessary to take the native structure of this protein as a target conforamtion, design a sequence to fit the target conformation, and fold this sequence. This requires the incorporation of side chains into the lattice model, and the recent work by Skolnick and Kolinsky (1993) demonstrated the feasibility of such an endeavor. We are currently working along these lines.

ACKNOWLEDGMENT

We thank Alexander Grosberg, Martin Karplus, Peter Leopold, Oleg Ptitsyn, and Andrej Sali for interesting and useful discussions. Graphic program ASGL by Andrej Sali was used to generate some of the plots.

REFERENCES

Baum, J., Dobson, C. M., Evans, P. A., & Hanly, C. (1989) *Biochemistry 28*, 7–13.

Baumgartner, A. (1984) *Annu. Rev. Phys. Chem. 35*, 419–435.

Bryngelson, J. D., & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 7524–7528.

Bryngelson, J. D., & Wolynes, P. G. (1989) *J. Phys. Chem. 93*, 6902–6915.

Briggs, M., & Roder, H. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 2017–2021.

Bycroft, M., Matouschek, A., Kellis, A., Jr., Serrano, L., & Fersht, A. R. (1990) *Nature 346*, 488–490.

Camacho, C. J., & Thirumalai, D. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 6369–6372.

Chaffotte, A., Guillou, Y., Delepierre, M., Hinz, H.-J., & Goldberg M. (1991) *Biochemistry 30*, 8067.

Covell, D., & Jernigan, R. (1990) *Biochemistry 29*, 3287–3294.

Creighton, T. (1992) *Proteins. Structure and Molecular Properties*, W. H. Freeman & Co., New York.

Elove, G., Chaffotte, A., Roder, H., & Goldberg, M. (1992) *Biochemistry 31*, 6876–6883.

Finkelstein, A. V., & Shakhnovich, E. I. (1989) *Biopolymers 28*, 1668–1694.

Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Ya. (1993) *FEBS Lett. 325*, 23–28.

Gittis, A. G., Stites, W. E., & Lattman, E. E. (1993) *J. Mol. Biol. 232*, 718–724.

Goldstein, R., Luthey-Schulten, Z. A., & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 4918–4922.

Hilhorst, H. J., & Deutch, J. M. (1975) *J. Chem. Phys. 63*, 5153–5161.

Hughson, F., Wright, P., & Baldwin, R. (1990) *Science 249*, 1544–1548.

Jeng, M. F., Englander, W., Elove, G., Wand, A., & Roder, H. (1990) *Biochemistry 29*, 10433.

Jennings, P., & Wright, P. (1993) *Science 262*, 892–896.

Karplus, M., & Weaver, D. (1976) *Nature 160*, 404–406.

Karplus, M., & Shakhnovich, E. (1992) in *Protein Folding* (Creighton, T. E., Ed.) pp 127–195, W. H. Freeman and Company, New York.

Kim, P., & Baldwin, R. (1982) *Annu. Rev. Biochem. 51*, 459–489.

Kolinski, A., & Skolnick, J. (1993) *J. Chem. Phys. 98*, 7420–7433.

Lau, K. F., & Dill, K. A. (1990) *Macromolecules 22*, 3986–3997.

Leopold, P. E., Montal, M., & Onuchic, J. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 8721–8725.

Levinthal, C. (1969) in *Mossbauer Spectroscopy of Biological Systems. Proceedings of a Meeting Held at Allerton House, Monticello, IL* (Debrunner, P., Tsibris, J.-C., & Munck, E., Eds.) pp 22–24, University of Illinois Press, Urbana, IL.

Lifshitz, E. M., & Pitaevskii, L. P. (1981) *Physical Kinetics*, Pergamon, Oxford, U.K.

Lifshitz, I. M., Grosberg, A. Yu., & Khohlov, A. R. (1978) *Rev. Mod. Phys. 50*, 683–713.

Matouschek, A., Kellis, J., Jr., Serrano, L., Bycroft, M., & Fersht, A. R. (1990) *Nature 346*, 440–445.

Matouschek, A., Serrano, L., & Fersht, A. R. (1992) *J. Mol. Biol. 224*, 819–835.

Mault, J., & Unger, R. (1991) *Biochemistry 30*, 3816–3824.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., & Teller, E. (1953) *J. Chem. Phys. 21*, 1087–1092.

Miller, R., Danko, C., Fasolka, M. J., Balazs, A. C., Chan, H. S., & Dill, K. A. (1992) *J. Chem. Phys. 96*, 768–780.

Miranker, A., Radford, S., Karplus, M., & Dobson, C. (1991) *Nature 349*, 633–636.

Miyazawa, S., & Jernigan, R. (1985) *Macromolecules 18*, 534–552.

O'Toole, E. M., & Panagiotoupoulos, A. Z. (1992) *J. Chem. Phys. 97*, 8644–8645.

O'Toole, E. M., & Panagiotoupoulos, A. Z. (1993) *J. Chem. Phys. 93*, 3185–3190.

Peng, S., & Kim, P. (1993) *Biochemistry* (in press).

Privalov, P. L. (1979) *Adv. Protein Chem. 33*, 167–241.

Privalov, P. L. (1992) In *Protein Folding* (Creighton, T. E., Ed.) pp 127–195, W. H. Freeman and Company, New York.

Ptitsyn, O. B. (1987) *J. Protein Chem. 6*, 273–293.

Ptitsyn, O. B. (1992) in *Protein Folding* (Creighton, T. E., Ed.) Chapter 6, pp 243–300, W. H. Freeman and Company, New York.

Ptitsyn, O. B., Pain, R., Semisotnov, G., Zerovnik, E., & Razglyaev, O. (1990) *FEBS Lett. 262*, 20–24.

Radford, S., Dobson, C., & Evans, P. (1992) *Nature 358*, 302–307.

Rey, J., & Skolnick, J. (1991) *Chem. Phys. 158*, 199.

Rooman, M. J., & Wodak, S. J. (1992) *Biochemistry 31*, 10239–10249.

Rooman, M. J., Kocher, J.-P., & Wodak, S. J. (1992) *Biochemistry 31*, 10226–10238.

Sali, A., Shakhnovich, E. I., & Karplus, M. (1994a) *J. Mol. Biol. 3*, 1614–1636.

Sali, A., Shakhnovich, E. I. & Karplus, M. (1994b) *Nature 369*, 248–251.

Segawa, S., & Sugihara, M. (1984) *Biopolymers 23*, 2473–2488.

Shakhnovich, E. I. (1994a) *Phys. Rev. Lett. 72*, 3907–3910.

Shakhnovich, E. I. (1994b) in *Protein Structure by Distance Analysis* (Bohr, H., & Brunack, S., Eds.) IOS Press, Amsterdam.

Shakhnovich, E. I., & Finkelstein, A. V. (1982) *Dolk. Akad. Nauk SSSR 243*, 1247–1251.

Shakhnovich, E. I., & Finkelstein, A. V. (1989) *Biopolymers 28*, 1667–1681.

Shakhnovich, E. I., & Gutin, A. M. (1989a) *Biophys. Chem. 34*, 187–199.

Shakhnovich, E. I., & Gutin, A. M. (1989b) *J. Phys. A22*, 1647.

Shakhnovich, E. I., & Gutin, A. M. (1990a) *J. Chem. Phys. 93*, 5967–5971.

Shakhnovich, E. I., & Gutin, A. M. (1990b) *Nature 346*, 773–775.

Shakhnovich, E. I., & Gutin, A. M. (1993a) *Proc. Natl. Acad. Sci. U.S.A. 90*, 7195–7199.

Shakhnovich, E. I., & Gutin, A. M. (1993b) *Protein Eng. 6*, 793–800.

Shankhnovich, E. I., Farztdinov, G. M., Gutin, A. M., & Karplus, M. (1991) *Phys. Rev. Lett. 67*, 1665–1667.

Shin, H. C., Merutka, G., Waltho, J. P., Tennant, L., Dyson, H., & Wright, P. (1993) *Biochemistry 32*, 6356–6366.

Skolnick, J., & Kolinski, A. (1990a) *J. Mol. Biol. 212*, 787–817.

Skolnick, J., & Kolinski, A. (1990b) *Science 250*, 1121–1125.

Skolnick, J., & Kolinski, A. (1991) *J. Mol. Biol. 221*, 499–531.

Sykes, M. (1963) *J. Chem. Phys. 39*, 410.

Tsong, T. Y., Baldwin, R., & McPie, P. (1972) *J. Mol. Biol. 63*, 453.

Ueda, Y., Taketomi, H., & Go, N. (1978) *Biopolymers 17*, 1531–1548.

Uversky, V., Semisotnov, G., Pain, R., & Ptitsyn, O. (1992) *FEBS Lett. 314*, 89–92.

Verdier, P. H. (1973) *J. Chem. Phys. 59*, 6119–6126.

Waltho, J. P., Feher, V. A., Merutka, G., Dyson, H., & Wright, P. (1993) *Biochemistry 32*, 6337–6355.

Wetlaufer, D. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697–701.

Williams, D., Harding, M., & Woolfson, D. (1991) *Biochemistry 30*, 3120–3128.

Wilson, C., & Doniach, S., (1989) *Proteins: Struct., Funct., Genet. 6*, 193–209.