Is Burst Hydrophobic Collapse Necessary for Protein Folding?†

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

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

Received August 25, 1994; Revised Manuscript Received November 30, 1994®

ABSTRACT: Folding of the lattice model of proteins is studied using Monte Carlo simulation. The amino acid sequence is designed to have a pronounced energy minimum for a given target (native) conformation. Our simulations reveal two possible scenarios. When the overall attraction between residues dominates, we find that folding to the native conformation is preceded by a rapid collapse into a burst intermediate which is a compact but structureless globule. Then, after a much longer time, an all-or-none transition from the globule to the native conformation occurs. In contrast, when the overall attraction is not strong, we do not observe a burst collapse stage. Instead, we find an all-or-none transition directly from the coil to the native conformation. Both scenarios yield comparable rates of folding. On the basis of these findings we discuss the role of intermediates in thermodynamics and kinetics of protein folding.

The famous Levinthal argument about the impossibility of finding a protein's structure by random search (Levinthal, 1969) stimulated the development of many phenomenological models aimed at explaining how a protein can fold, avoiding a random search over its huge conformational space (Karplus & Weaver, 1979; Weaver & Karplus, 1994; Kim & Baldwin, 1982; Harrison & Durbin, 1985; Wetlaufer, 1973). The underlying idea for most phenomenological models of folding concerned the existence of some pathway, different in different models, with several stages and intermediates. This introduced hierarchical elements to folding in which certain degrees of freedom become "frozen" so that the remaining search is feasible, since it required scanning over a much smaller number of conformations. A well-known example of this kind is a secondary structure framework or the related, diffusion−collision model (Kim & Baldwin, 1982; Karplus & Weaver, 1979; Weaver & Karplus, 1994). These models postulate that certain elements of native structure, microdomains (usually associated with elements of native secondary structure), are formed very fast, and in subsequent folding they move as a whole and associate via, e.g., the diffusion−collision mechanism.

A similar way to decrease the effective number of conformations searched is to use a two-stage kinetics, the first stage being compactization into structureless globule. At the second stage the search for the native conformation takes place among only compact globular conformations whose number is much less than the total number of conformations (Orland *et al.*, 1985; Dill, 1985). Therefore, compactization should facilitate the search for the native conformation.

The compact state without specific structure (judged by lack of specific protection against H−D exchange in most parts of a protein) was indeed observed as a "burst" intermediate in a number of proteins: hen egg lysozyme (Redford *et al.*, 1992; Itzhaki *et al.*, 1994), cytochrome *c* (Chaffotte *et al.*, 1991), and a tryptophan-containing mutant of ubiquitin at 25 °C (Khorasanizadeh *et al.*, 1993). This burst intermediate is a compact conformation with few specific contacts and some secondary structure detected by CD which is likely to be fluctuating without any specific localization. The evidence for that is absence of specific protection of the amides in the burst intermediate.

The two-stage kinetics was also observed in computer simulations of protein folding (Shakhnovich *et al.*, 1991; Miller *et al.*, 1992; Honeycutt & Thirumalai, 1992; Camacho & Thirumalai, 1993; Fukugita *et al.*, 1993; Sali *et al.*, 1994a; Socci & Onuchic, 1994). It was argued that burst compactization plays an important role in making the subsequent random search for a transition state feasible. However, it was noted (Sali *et al.*, 1994) that such mechanism is likely to be applicable to folding of relatively short, prebiological peptides with almost random sequences. As chain length N increases to realistic sizes, random sequences fail to fold (Shakhnovich, 1994). This is due to the fact that for longer sequences the number of compact conformations is still exponentially large in N.

The point of view that Levinthal paradox can be resolved only by sequential mechanism involving intermediates enjoyed indirect support from experiment since in many studies compact intermediates were found indeed. However, the generality of intermediates was questioned in a recent study of Sosnick *et al.* (1994), who showed that folding of cyt *c* at pH 5 differs dramatically from the scenario described in earlier studies and represents essentially the two-state kinetics where all observed properties become native-like simultaneously and fast. And cyt *c* is not unique. Another protein, chymotrypsin inhibitor 2, folds kinetically via a two-state mechanism showing no intermediates (Jackson & Fersht, 1991). Moreover, in a detailed study of ubiquitin folding, Khorasanizadeh *et al.* (1993) showed that, depending on the final concentration of the denaturing agent after dilution, the same protein may or may not have folding intermediates. Sosnick *et al.* (1994) suggested to distinguish "artificial or adventitious" barriers for folding (such as slow heme ligation in cyt *c* in their case) from the barriers intrinsic for folding. In the same way, one can question the role of folding intermediates because indeed there is a number of experimental examples where folding apparently proceeds

† Supported by the Packard Foundation.

* Address correspondence to this author (telephone 617-495-4130; FAX 617-496-5948; Internet: eugene@diamond.harvard.edu).

® Abstract published in *Advance ACS Abstracts,* February 15, 1995.

0006-2960/95/0434-3066$09.00/0

without them. Moreover, one can imagine that burst compactization may result in slowing of certain chain motions because parts of the chain move in dense environment.

Thus, questions arise: Is compactization a *necessary* separate step of folding or can it take place simultaneously with the native structure formation? Does nonspecific "hydrophobic collapse" facilitate folding?

It is difficult to answer these questions based exclusively on the analysis of experimental data because in most cases such analysis requires certain assumptions (such as linear dependence of free energy of the unfolded state and transition state on denaturant concentration), and there may always be questions concerning the time resolution of the kinetic apparatus. The computational approach to this question seems very useful as it does not have limitations of time resolution and allows the eventual tracing of any intermediate conformation, and therefore one can be sure that nothing is "hidden" in numeric experiments. The two crucial requirements to the computational model is that model proteins should possess an astronomically large number of conformations and that simulations should reach the unique native conformation and subsequently stay in the native state, which includes native conformation and fluctuations around it. Such fluctuations are always present at finite temperature.

The only available computational models in which both requirements are fulfilled are simple lattice models of proteins (Skolnick & Kolinski, 1991; Shakhnovich *et al.*, 1991; Miller *et al.*, 1992; Camacho & Thirumalai, 1993; Sali *et al.*, 1994a; Socci & Onuchic, 1994). Of course, these are idealized models in which the virtue of folding is achieved by significant simplifications such as presentation of amino acid residues as structureless beads. Hence, it can tackle only such problems where detailed stereochemistry and packing of side chains may be not relevant. The most important of these are aspects of the protein folding problem concerned with formation of native chain topology, which is believed to occur during the fast stages of folding. In this study we are interested in the role of burst intermediates which may be formed at the fastest stages of folding. Therefore, we believe that the simple lattice model may be adequate to tackle this aspect of the protein folding problem.

In the present work we study the role of burst intermediates using a simple lattice model of folding. We show that hydrophobic collapse is not necessary for folding. Depending on solvent conditions or amino acid composition of a protein (*viz.*, whether monomers on average attract each other or not), burst collapse may precede folding to the native state, or compactization may proceed simultaneously with folding to the native state. In the latter case the resolution of the Levinthal paradox in our model system does not require that intermediates be formed in the process of folding.

THE MODEL

We consider a protein chain positioned on a cubic lattice. An amino acid residue is presented by a structureless monomer occupying a lattice site. Residues connected by a covalent bond must occupy neighboring lattice sites. Two or more monomers cannot occupy the same site.

The energy of a conformation is a sum of energies of pairwise contacts between monomers:

$$E = \sum_{1 \le i < j \le N} U(\xi_i,\xi_j)\Delta_{ij} \qquad (1)$$

where $\Delta_{ij} = 1$ if monomers i and j are lattice neighbors and $\Delta_{ij} = 0$ otherwise. ξ_i defines the specific amino acid residue in position i along a chain. $U(\xi,\eta)$ is the interaction energy between the amino acid residues ξ and η.

Different aspects of the Hamiltonian eq 1 are important when protein molecules are in different thermodynamic states. Possible states of a protein molecule were characterized in the mean-field theory of heteropolymers (Shakhnovich & Gutin, 1989); these results were later confirmed by exact calculations for short chains with exhaustively enumerated conformations (Dinner *et al.*, 1994; Kantor & Kardar, 1994). When the chain is in the state where it is compact but disordered, without unique structure, its thermodynamics are described by average properties of the potential (Shakhnovich & Gutin, 1989). This can be understood if we note that multitudes of nonsimilar conformations are equally probable in the denatured state so that each individual interaction does not persist for a long time. This leads to effective "averaging out" of individual interactions. Correspondingly, since the disordered state does not have a specific conformation, its only relevant thermodynamic characteristic is density, which is determined by average interactions: if it corresponds to attraction, then the chain is compact in its disordered state, and it is random coil when the average interaction is repulsive. In the opposite case, when the chain has unique structure, each interaction persists indefinitely, and each individual interaction strength is important.

Hence it us useful to present the Hamiltonian in eq 1 in the form where average interaction is singled out explicitly:

$$E = \sum_{1 \le i < j \le N} (B_0 + \tilde{U}(\xi_i,\xi_j))\Delta_{ij} \qquad (2)$$

where

$$B_0 = \frac{2}{N(N-1)} \sum_{i<j}^{N} U(\xi_i,\xi_j)$$

is average interaction energy (averaged over all possible contacts) and $\tilde{U}(\xi_i,\xi_j)$ is the interaction energy taken from Table VI of the work of Miyazawa and Jernigan (1985).

Presentation of the interaction energy in the form of eq 2 allows separation of specific from nonspecific contributions. This separation is helpful as it allows taking into account the influence of the solvent or mutations on the native state and disordered denatured state separately. Indeed, the disordered conformation is very susceptible to solvent conditions, as any polymer below or at the Θ point (Lifshitz *et al.*, 1978; De Gennes, 1979; Finkelstein & Shakhnovich, 1989). Opposite to it, the native state is robust and does not change very much in a range of solvent conditions or under influence of some mutations.

Therefore, semiqualitatively, the effect of changing solvent composition or making mutations can be modeled through changing of B_0. Though in reality altering solvent composition does not change uniformly all parameters $U(\xi,\eta)$, its most important influence is that it varies density of the denatured state, while the native structure remains unchanged in a range of solvent conditions. The kinetic counterpart of

the compact denatured state is a burst intermediate; therefore, stability of this state is a very important determinant of its existence in kinetics. As was pointed out previously, all thermodynamic characteristics of the denatured state are determined by B_0; this is why this parameter is very important for the present study. For example, influence of solvent or mutations on the density and stability of the compact intermediate can be adequately described through the change in the average value of B_0.

The stability of this state is of key importance for the present discussion, which focuses on the condition of existence of the burst compact kinetic intermediate.

We should also note that parameters of Miyazawa and Jernigan (1985) were obtained from protein statistics in quasichemical approximation, and therefore their meaning is only relative; there is no intrinsic energy scale in the procedure of derivation of energy parameters from probabilities of occurrence of different contacts in proteins (Finkelstein *et al.*, 1992). Therefore, we scale parameters by multiplying each $\tilde{U}(\xi_i,\xi_j)$ by a constant factor to make the standard deviation

$$\langle \tilde{U}^2 \rangle = \frac{2}{N(N-1)} \sum_{i<j}^{N} \tilde{U}^2(\xi_i,\xi_j) = 1$$

We showed earlier (Sali *et al.*, 1994a; Shakhnovich, 1994; Abkevich *et al.*, 1994a) that the necessary and sufficient condition for sequences to fold in this model is that the native state be a pronounced energy minimum for this sequence. Hence, the sequence design was aimed at generation of such sequences. The detailed discussion of the design algorithm, Monte Carlo optimization in sequence space, was published earlier (Shakhnovich & Gutin, 1993a,b). The design approach requires first picking a target conformation for which sequences should be designed so that this target conformation will be native for them (i.e., stable and kinetically accessible). We used the conformation shown in Figure 1 as the target one; this native conformation is the same as was used in our previous studies (Abkevich *et al.*, 1994a,b).

Unlike the folding problem, the solution of the design problem is not unique, and the design algorithm can generate thousands of nonhomologous sequences having low energy in the target conformation (Figure 1). We tested a number of them (more than 50); each folded to the target conformation. For our study we used the sequence shown on Figure 1b; previous results (Abkevich *et al.*, 1994b) suggested that key features of the folding mechanism were independent of the sequence chosen, provided that it is a folding sequence for its native conformation.

Folding simulations are made using the standard Monte Carlo method for polymers on a cubic lattice. The move set includes corner flips and crankshaft motions. The detailed discussion of advantages and caveats of the lattice Monte Carlo approach to study folding has been discussed (Skolnick & Kolinski, 1991; Rey & Skolnick, 1991; Sali *et al.*, 1994b).

Simulations begin with different random coil conformations in different runs. In each run the chain folded into the target native conformation. We made several long runs in order to make sure that no conformations with energy lower than that of the native state are encountered. This was indeed the case which made us believe that the target conformation

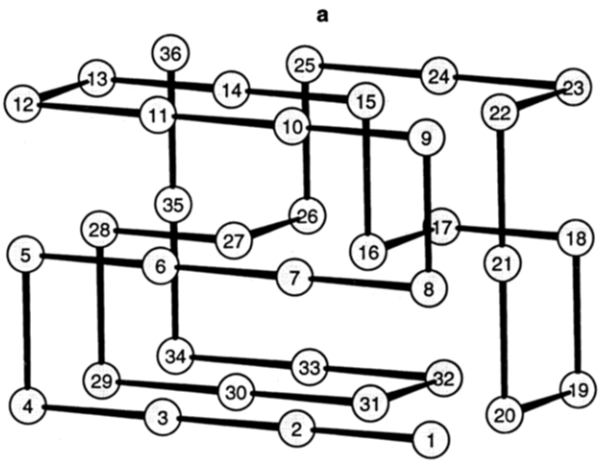

FIGURE 1: (a) A maximally compact conformation of a 36-mer on a cubic lattice and (b) a sequence for which this conformation is native.

shown in Figure 1a is the global energy minimum for the sequence shown on Figure 1b.

RESULTS

Two different regimes of folding were found. In order to illustrate these differences in the most pronounced way, two different sets of parameters were chosen for the detailed study. The first set corresponds to overall attraction between residues having $B = -1.0$ (see eq 1); simulations were run at temperature $T = 1.25$. We will refer to this set as the attraction case. The second set is $B = 0.4$ and $T = 0.65$. In this case two randomly taken residues are likely to repel each other; hence, we will refer to this set as to the repulsion case. In both cases the temperature was chosen to satisfy the two following conditions. The first goal is to make the native state thermodynamically stable. This condition imposes an upper limit for the temperature interval. The second condition requires relatively rapid folding into the native conformation. This limits the temperature from below, since it is known that folding slows down at low temperatures (Bryngelson & Wolynes, 1989; Miller *et al.*, 1992; Sali *et al.*, 1994a; Socci & Onuchic, 1994; Guo & Thirumalai, 1994).

For both sets of parameters we performed 1000 MC folding runs starting from different randomly generated coil conformations. In every 10 000 MC steps we monitored the number of all contacts C and the number of native contacts N in a current conformation. Here by native we mean all the particular contacts present in the native conformation (Figure 1a). The number of all contacts C can be treated as a measure of compactness of a chain. It can vary from 0 for a fully open conformation to 40 for a maximally compact conformation. The number of native contacts N shows how close a given conformation is to the native one. This number varies from 0 to 40 as well because the native conformation in our model is maximally compact (Figure 1a).

Figure 2 shows MC−time dependence of the number of native contacts $\bar{N}$ and the total number of contacts $\bar{C}$ averaged over 1000 runs. Such averaging over a large number of runs corresponds to ensemble averaging over individual molecules in solution. It can be seen that the averaged number of the

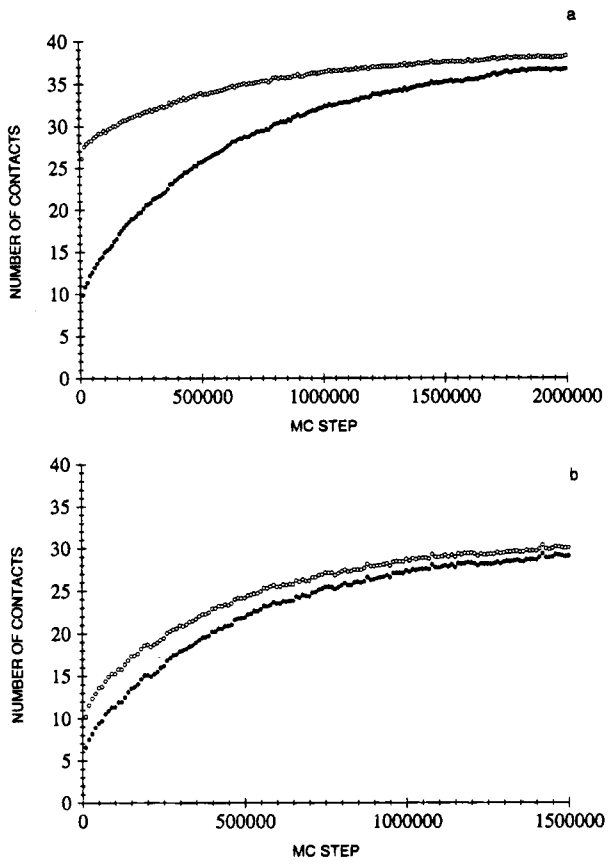

FIGURE 2: Monte Carlo evolution of the number of contacts averaged over 1000 runs starting from random coil conformations (a) in the attraction case with $B = -1.0$ and $T = 1.25$ and (b) in the repulsion case with $B = 0.4$ and $T = 0.65$. Filled circles correspond to the native contacts (top curves), and open circles correspond to all contacts (bottom curves).

native contacts $\bar{N}$ grows monotonically with MC−time and reaches values close to 40 in about 10^6 MC steps. This provides an estimate for the folding time. Though the MC−time dependence for the averaged number of native contacts $\bar{N}$ is similar in both the attraction and the repulsion cases, the time course for the averaged number of all contacts $\bar{C}$ (related to chain compactness) exhibits very different behavior. In the attraction case, in an average of 10^4 MC steps the chain acquires about 26 contacts, only 10 of them being native. After that the average number of all contacts $\bar{C}$ changes marginally. Therefore, there is a 2 order of magnitude difference between folding and compactization rate in this case.

In contrast, in the repulsion case after 10^4 MC steps on average the chain acquires only about 10 contacts, 7 of them being native. Moreover, the average number of all contacts $\bar{C}$ is close to the average (over all 1000 runs) number of the native contacts $\bar{N}$ all the time during folding. This suggests that in the repulsion case compactization and folding to the native state proceed simultaneously.

Figure 2 gives only the averaged (over the 1000 runs) characteristics of the chain. It does not say much about the behavior of the chain in a single run. More information can be derived from distributions of the number of native and all contacts at a given MC step. The corresponding histograms are shown in Figures 3 and 4 for the attraction case and in Figures 5 and 6 for the repulsion case. In both cases data are collected over 1000 runs at three different MC steps: (a) long before folding is completed, (b) when folding is about halfway completed, and (c) when folding is almost completed, which corresponds to equilibrium.

Let us first consider the attraction case (Figures 3 and 4). At the early stage of folding (4×10^4 MC steps) the distributions of N and C are monomodal and have very pronounced maxima (Figures 3a and 4a). This means that in all runs the chains form the same (burst) intermediate state. The number of the native contacts N in this state fluctuates around 10, and the number of all contacts C fluctuates around 29. Thus, this state is a compact non-native globule. At the later stage of folding (5×10^5 MC steps) histograms for the distribution of native and total contacts over all runs become bimodal: One can distinguish clearly two different maxima on both N and C histograms (Figures 3b and 4b). For native contacts the maxima are approximately at $N = 10$ and at $N = 40$. For total contacts the maxima are located around $C = 29$ and $C = 40$. The presence of two maxima on the histograms suggests that there are two different states separated by a first-order intramolecular transition.[1]

One of them, with N and C fluctuating around 40, is, obviously, the native state. The other state is similar to the state which had been formed at the early stage of folding, having on average about the same compactness and about the same number of the native contacts. At last, when the folding is almost completed (2×10^6 MC steps), the distributions are monomodal again (Figures 3c and 4c) with maximum corresponding to the native state.

Summarizing these data, we arrive at the following folding scenario in the attraction case. Starting from a coil conformation the chain very rapidly ($\sim 10^4$ MC steps) collapses into a compact ($C \approx 29$) non-native ($N \approx 10$) intermediate. Then an all-or-none transition from the intermediate to the native conformation takes place over much longer time ($\sim 10^6$ MC steps). This is just the typical scenario which was observed in a number of previous simulations (Shakhnovich *et al.*, 1991; Miller *et al.*, 1992; Honeycutt & Thirumalai, 1992; Camacho & Thirumalai, 1993; Fukugita *et al.*, 1993; Sali *et al.*, 1994a; Socci & Onuchic, 1994).

Now let us consider the repulsion case (Figures 5 and 6). At the early stage of folding (5×10^4 MC steps) the distributions of N and C are monomodal and have very broad maxima centered at $N = 7$ and $C = 12$. Such a small number of all contacts corresponds to a noncompact state which should be associated with a random coil. It should be mentioned that on average more than half of all contacts are native in this state. At a later stage of folding (3×10^5 MC steps) two different maxima can be distinguished clearly only on the histogram for the distribution of the number of native contacts, N (Figure 5b). These maxima are located at $N = 7$ and at $N = 33$. Though the distribution of C does not show clearly two maxima (Figure 6b), it is clear that this distribution is not monomodal. Moreover, the distribution for the total number of contacts, C, can be roughly approximated as a superposition of two broad maxima, one of which is centered at $C = 12$ and the other at $C = 33$. The

[1] To avoid confusion, we should note that we are speaking about two different *states*, defined thermodynamically as *set of conformations* corresponding to a free energy minimum. The existence of two states separated by a free energy barrier is a signature of an intramolecular first-order transition (Karplus & Shakhnovich, 1992). This is equivalent to the concept of the "all-or-none" transition accepted in protein chemistry to describe cooperativity of protein denaturation (Creighton, 1992).

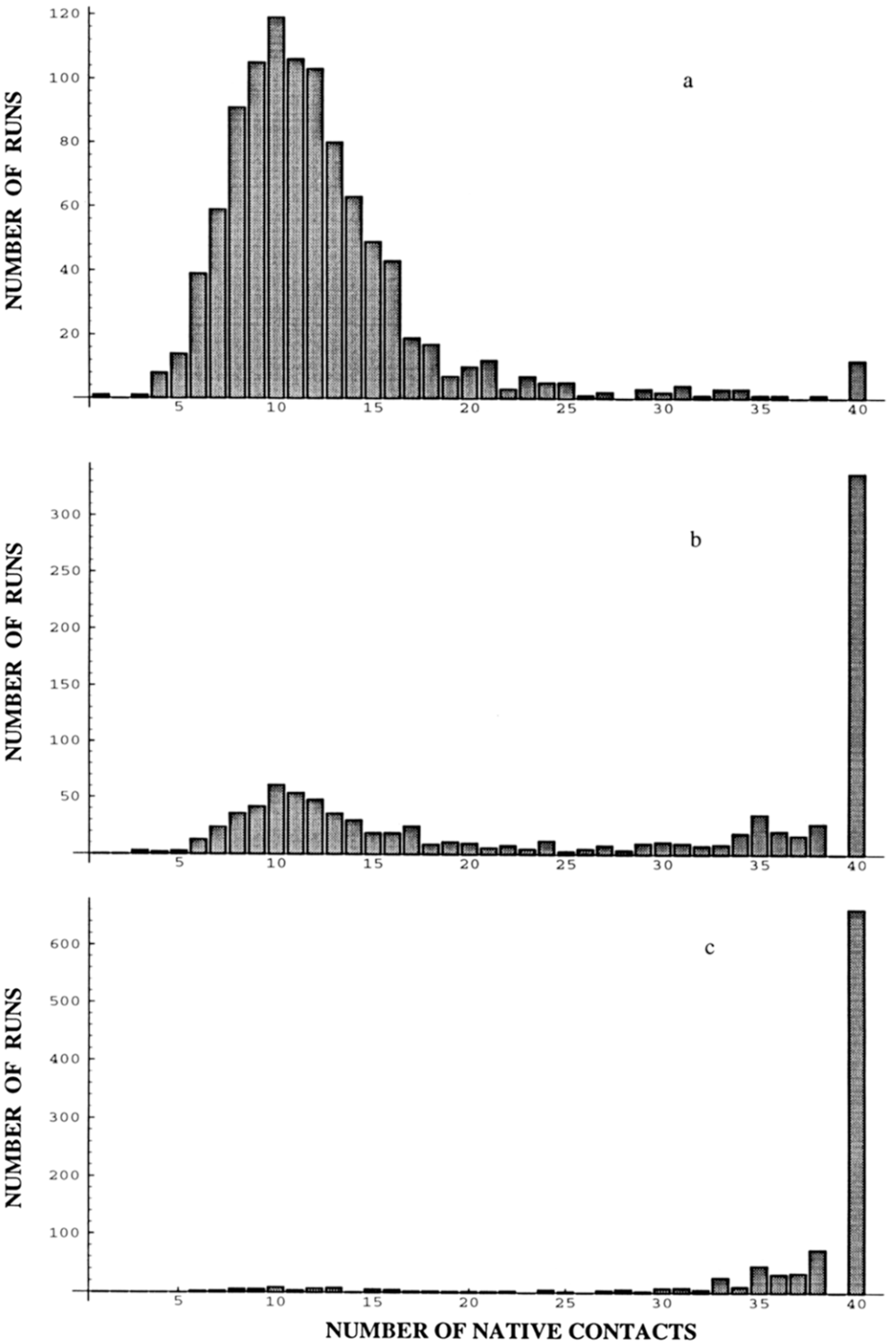

FIGURE 3: Distribution of the number of native contacts over 1000 runs at different stages of folding in the attraction case: (a) after 4×10^4 MC steps, (b) after 5×10^5 steps, and (c) after 2×10^6 steps.

presence of two maxima suggests, as usual, that there are two different states separated by the free energy barrier. One of them, corresponding to N and C both fluctuating around 33, should be identified as the basin of attraction of the native state. The other state is very close to the coil-like state found at the early stage of the folding. At last, when folding is almost completed (1.5×10^6 MC steps), the distributions are basically monomodal with maxima corresponding to the native state.

These results can be summarized to describe the folding scenario in the repulsion case. There is no evidence for a compact non-native intermediate in this case. Folding into the native state proceeds directly from the coil-like state to the native one as an all-or-none transition.

DISCUSSION

Using a simple lattice model, we showed that, depending on the value of the average interaction between monomers (roughly modeling the effect of solvent or overall amino acid composition of a protein), two different regimes of folding are possible. When strong attraction between monomers dominates, two-stage kinetics is observed. The first stage is fast collapse into a compact non-native globule, and the second stage is a slow search for the native conformation

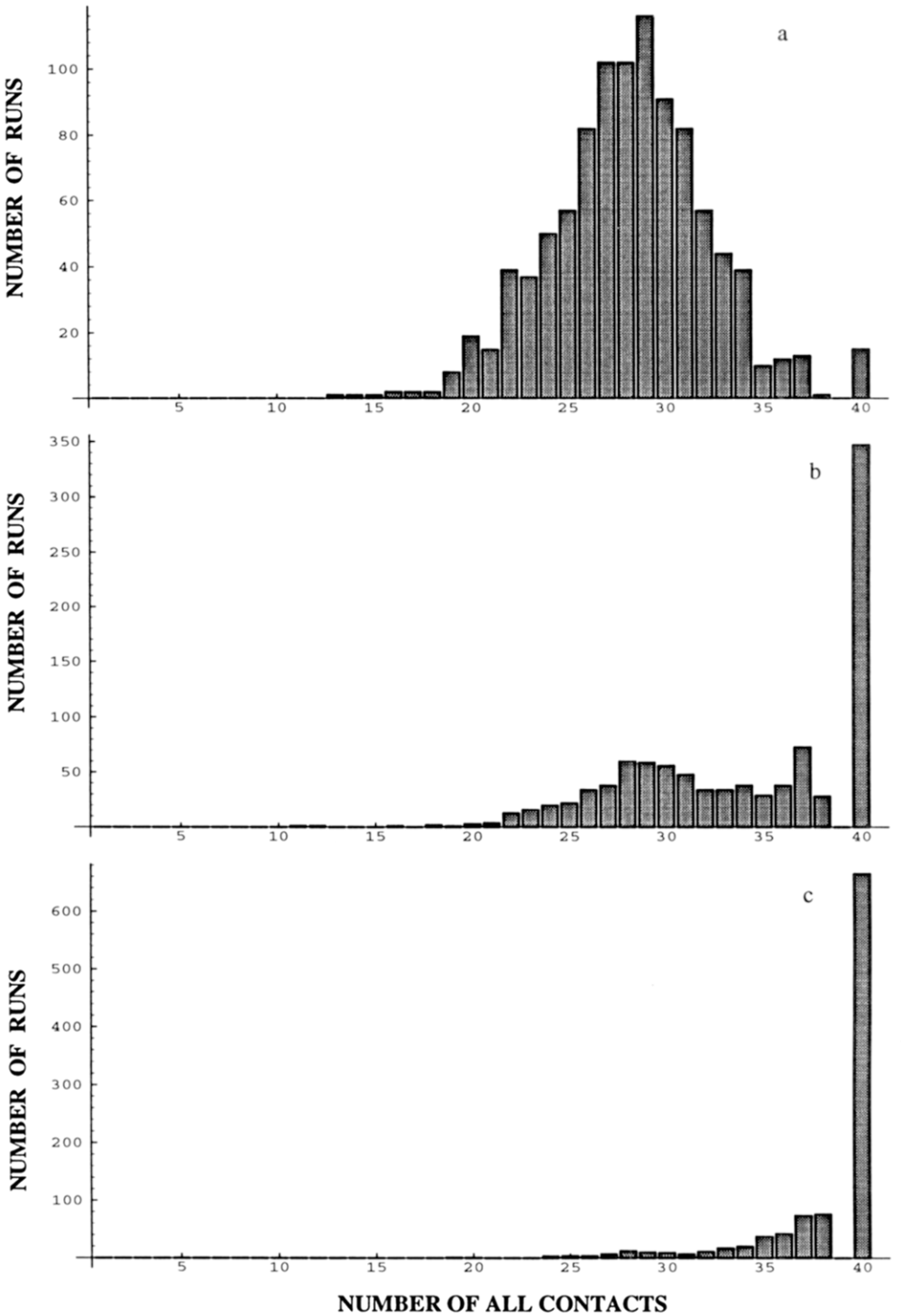

FIGURE 4: Distribution of the number of all contacts over 1000 runs at different stages of folding in the attraction case: (a) after 4×10^4 MC steps, (b) after 5×10^5 steps, and (c) after 2×10^6 steps.

among only compact ones. Folding into the native state from the compact non-native intermediate is an all-or-none transition. When there is average repulsion between monomers, compactization and the native structure formation proceed simultaneously as an all-or-none transition between random coil and native states.

It should be mentioned that for real proteins the average interaction between amino acid residues strongly depends on the concentration of denaturants, pH, and temperature. Therefore, two described regimes can be found for the same protein under different folding conditions; this happens for the Trp-containing mutant of ubiquitin (Khorasanizadeh *et al.*, 1993). Moreover, average interaction between residues depends on the amino acid composition of a protein as well. For example, proteins more hydrophobic in composition have stronger attraction between residues at the same conditions. Presence of disulfide bonds also effectively results in stronger attraction. Therefore, at the same folding conditions different proteins may follow different folding scenarios.

The conditions under which a particular scenario of folding takes place can be illustrated in a schematic free energy−density diagram (Figure 7). It can be seen that sequences which allow the possibility of folding from a coil directly to the native state must satisfy certain conditions. Indeed, as

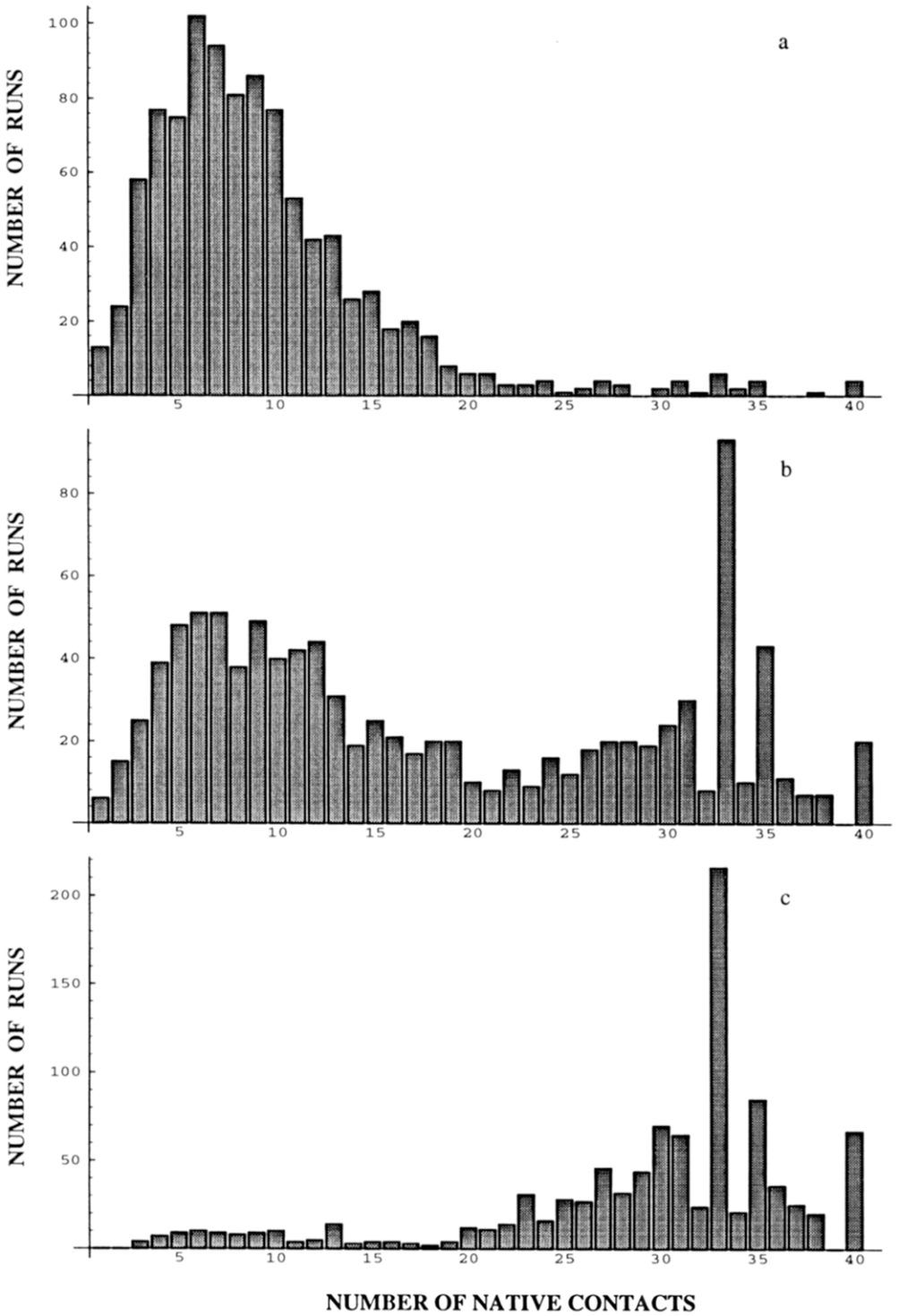

FIGURE 5: Same as Figure 3 but for the repulsion case.

Figure 7 indicates, the energy difference between compact denatured states and the native state must be sufficiently large to allow coil states to be "in between", i.e., have energy below the compact denatured states but above the native states. Moreover, this energy difference must be pronounced in order to make compact denatured states energetically unfavorable, suppressing their possible presence. This implies that the energy gap between the native and compact denatured states must be sufficiently large, i.e., that sequences must be strongly optimized (Goldstein *et al.*, 1992; Shakhnovich & Gutin, 1993a; Shakhnovich, 1994) in order to allow (at certain conditions) direct folding from a random coil to the native state.

Another explanation of that fact employs the following thermodynamic argument. Upon folding transition the folded state must dominate thermodynamically. That means that energetically favorable interactions in the native conformation must be sufficient to suppress the excess entropy of the unfolded state. In a case when the unfolded state is noncompact, this entropy is higher than in the case when the unfolded state is compact. Indeed, the number of compact conformations is $\sim e^N$ times less than the number of all conformations (Orland *et al.*, 1985; Dill, 1985; Camacho & Thirumalai, 1993b). Elementary statistical mechanics suggests that the entropic contribution to the free energy of the coil constitutes approximately an extra $-k_BT$

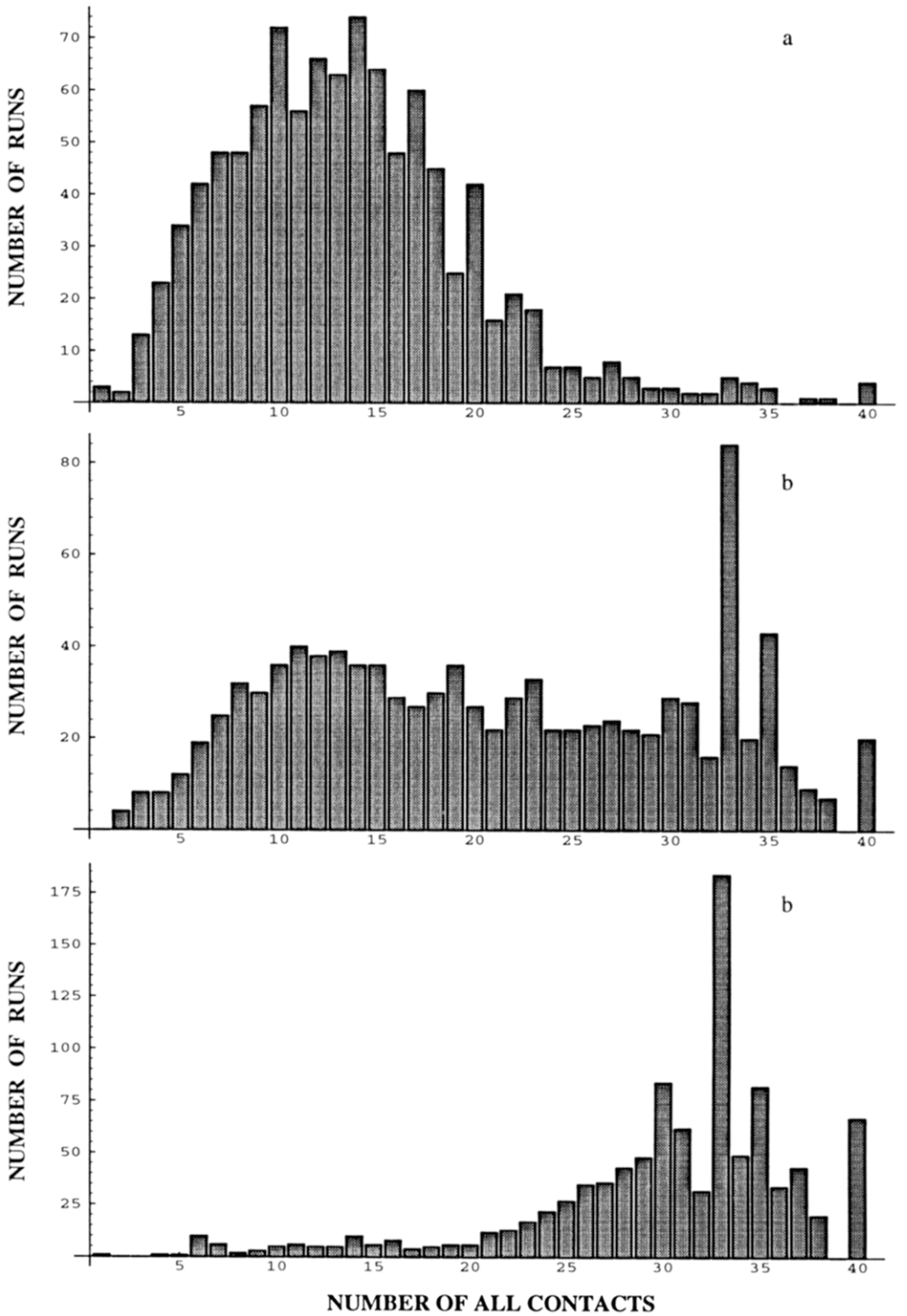

FIGURE 6: Same as Figure 4 but for the repulsion case.

per residue compared to that of the compact denatured state (k_B is Boltzmann's constant). This extra entropy must be suppressed by better energetic optimization of sequences to provide more favorable contacts in the native state, in order to make folding from the random coil effective.

This analysis has a number of implications. First, it follows that only highly stable proteins may, under certain conditions, fold directly from a random coil. Second, previous simulations with short quasirandom sequences (Shakhnovich *et al.*, 1991; Miller *et al.*, 1992; Camacho & Thirumalai, 1993; Sali *et al.*, 1994a,b) revealed that, as a rule, burst compactization precedes folding. We see now that this is due to the fact that these sequences, being weakly optimized (having relatively small energy gaps) have free energy landscapes as shown in Figure 7a. This implies that a pronounced average dominant attraction between monomers is a thermodynamic prerequisite for stability of folded conformations for such sequences. It is well-known, both from analytical theory (Grosberg & Shakhnovich, 1986; Shakhnovich & Gutin, 1989a) and simulations (Shakhnovich *et al.*, 1991; Camacho & Thirumalai, 1993; Socci & Onuchic, 1994) that collapse transitions in heteropolymers do not require overcoming the free energy barrier (second-order transition). For this reason average attraction between residues gives rise to burst compactization in kinetics, following the simple mechanism of heteropolymer collapse

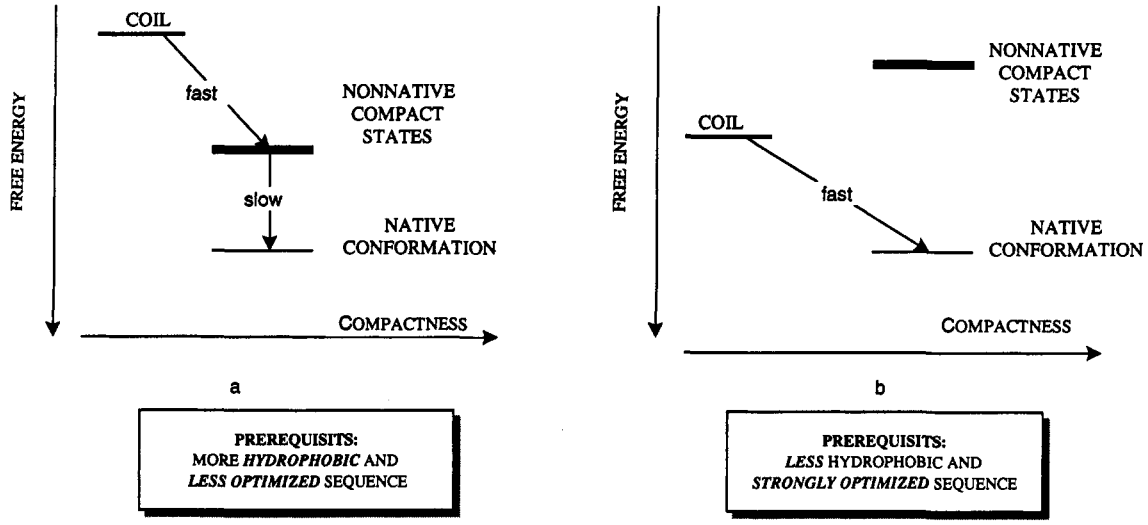

FIGURE 7: Schematic representation of the free energy landscape in the attraction case (a) and in the repulsion case (b).

(Socci & Onuchic, 1994). However, this is not the factor accelerating subsequent folding to the native state, as was shown in the present work.

Many experimental studies point out that, similar to fast collapse, a certain fraction of native or non-native secondary structure forms very fast as judged by far-UV CD kinetics (Kuwajima *et al.*, 1987; Gilmanshin & Ptitsyn, 1987; Goldberg *et al.*, 1990; Redford *et al.*, 1992; Elöve *et al.*, 1992). It is tempting to suggest that this is important for protein folding kinetics, especially in view of the fact that formation of secondary structure was assumed to play a crucial role in resolving the Levinthal paradox in many phenomenological models [reviews in Kim and Baldwin (1982) and Karplus and Shakhnovich (1992)]. However, the present study gives us some ground to question the kinetic importance of secondary structure in spite of the fact of its fast formation.

Indeed, we may suggest, in analogy with the case of compactization, that secondary structure plays an important *thermodynamic* role in proteins (and that is why it is one of the most invariant features of protein architecture). The reason why secondary structure may be important for thermodynamics can be found in analytical theories of protein folding (Bryngelson & Wolynes, 1987; Shakhnovich & Gutin, 1989a, 1990). Introduction of secondary structure makes the chain stiffer, eliminating many competing non-native conformations incompatible with the native-like secondary structure. As a direct consequence of thermodynamic stabilization of secondary structure, some elements of it may form fast, because their formation involves very small free energy barriers (Poland & Scheraga, 1970)—analogous to the compactization case discussed in this paper. However, as in the case of compactization, it well may be that such fast formation of secondary structure on the burst stage is not very important for subsequent folding *kinetics* to the native state. Recent calculations (Shakhnovich, 1994; Abkevich *et al.*, 1994a,b) provided numerous examples that the Levinthal paradox can be solved for folding of long lattice model chains without any involvement of secondary structure formation. Certainly the above arguments represent only a possible explanation of the role of secondary structure, and much more work, experimental and theoretical, is necessary to come to a definitive answer to this important question.

The important aspect of the present work should include a "reality check", i.e., whether the thermodynamic characteristics of model proteins resemble those of real proteins. Let us first consider the attraction case. First, we estimate the free energy of stabilization of the native state. This can be done by comparing the areas A_f, A_u under the native state peak and the denatured state peak in Figure 3c. Since the transition from the denatured state to the native proceeds through the main free energy barrier (corresponding to the number of the native contacts $N \approx 20$, on Figure 3c), we apply the equations of the two-state thermodynamics which give $A_f/A_u = \exp(-(G_f - G_u)/k_BT)$, where $G_{f,u}$ is the free energy of the folded and unfolded states, respectively. The straightforward calculation yields that $\Delta G = G_f - G_u = -2.5k_BT$ at $T = 1.25$. We have to extrapolate this number to "physiological conditions". To this end we note that the mid-transition for our lattice is at $T_f = 1.41$. (We define mid-transition temperature as that at which the thermodynamically averaged number of native contacts $\langle N \rangle_{th} = N_{max}/2 = 20$ for 36-mer with $N_{max} = 40$.) Since the transition is first-order-like, $\Delta G \sim (T - T_f)/T_f$ (Stanley, 1971; Karplus & Shakhnovich, 1992; Shakhnovich & Finkelstein, 1989). "Physiological conditions" for proteins correspond to temperature $T \approx 300$ K while denaturation temperature $T_f \approx 350$ *K*; i.e., for real proteins $(T - T_f)/T_f \approx 0.2$. This means that for our model proteins the "physiological temperature" would correspond to $T \approx 1.1$. The stability of our lattice model proteins at this temperature would be $\Delta G \approx 5k_BT$ for the 36-monomer chain. The 100-residue chain at these conditions would have $\Delta G \approx 15k_BT$, or $\approx$10 kcal/*m* if we take $T = 330$ K as a real world counterpart of $T = 1.25$ in our simulations. This free energy of stabilization is close to the one observed in real proteins (Privalov, 1979). We should note here that we have deliberately chosen the simulation temperature $T = 1.25$ for the attraction case to be somewhat high at $(T_f - T)/T_f = 0.1$ instead of 0.2 for real proteins at physiological conditions. The reason is that to get decent statistics, we have to run thousands of simulations, and therefore we need to make folding somewhat faster to make such computations feasible.

An analogous thermodynamic analysis can be done for the repulsion case. In order to achieve stabilization of the native state in the repulsion case, we had to decrease

temperature to $T = 0.65$. The reason is that in the repulsion case the denatured state has much higher entropy as it includes not only compact conformations but also a multitude of noncompact conformations, and lower temperature is needed to suppress configurational entropy of the denatured state in the present case. Even such decrease of temperature helps only partly since fluctuations around the native conformations are pronounced (Figure 5). In that case the most probable is not the native conformation but the set of conformations having 33 out of 40 native contacts differing from the native conformation by a detached "weak loop" forming the least favorable contacts with the remaining part of the structure.

This problem at low temperature uncovers the caveat of the present lattice model, which is high flexibility of lattice chains compared to real polypeptide chains. The important measure of a polymer flexibility is the number of conformations per residue. For free polypeptide chains this number was estimated by Dill (1985) to be 2.40. For free cubic lattice chains this number is 4.68 (De Gennes, 1979). Compactization decreases the number of conformations in the same way as chain stiffening, and therefore in the attraction case thermodynamics of lattice proteins are very close to thermodynamics of real proteins.

The alternative way of "stiffening" the chain is introduction of secondary structure. This will also decrease entropy of the denatured state, leading to enhanced stability of the native state. The implication of it would be that folding in the repulsion case is possible at higher temperature. The other expected result of stiffening the backbone by introduction of secondary structure is that fluctuations of the native state will be much less pronounced even in the repulsion case. In fact, compactization and formation of secondary structure may be thermodynamically equivalent, as was pointed out by Chan and Dill (1990).

Another new aspect which can be introduced by secondary structure is that for stiff chains compactization transition is cooperative while flexible chains undergo noncooperative compactization (Lifshitz *et al.*, 1978; Kolinski *et al.*, 1986). The cooperativity of compactization transition was studied in ubiquitin by Khoranizadeh *et al.*, (1993), where they plotted the amplitude of the burst phase vs final dilution and obtained a sigmoidal curve, which is highly suggestive of cooperative transition. This points out to important *thermodynamic* role of secondary structure. Future development of lattice models will include the possibility of secondary structure formation.

Finally, we would like to discuss how are our results are related to the Levinthal paradox. Its traditional formulation emphasizes the entropic cost of folding, focusing on the large number of conformations available to a polypeptide chain. Camacho and Thirumalai (1993b) suggested that the Levinthal paradox may be resolved by a three-stage procedure: at the first stage a chain undergoes overall compactization, at the second stage it falls into one of the minimum energy compact conformation (MECS), and at the third stage it searches for the native conformation over a relatively small number of MECS. This hypothesis (as well as many other similar models of folding) ignores the fact that though the number of relevant conformations may decrease, barriers between them grow (Chan & Dill, 1994), such that interconversion between these fewer conformations slows down dramatically. This shows no solution to the Levinthal paradox. Indeed, several analytical studies of kinetics of random systems, including simplified protein models (De Dominicis *et al.*, 1985; Kopper & Hilhorst, 1987; Bryngelson & Wolynes, 1989; Shakhnovich & Gutin, 1989b), demonstrated this property, showing that barriers between low-energy misfolded ("glassy") states may well be $\sim Nk_BT$, so that such states represent deep traps, escape from which is extremely slow. Therefore, arguments based on the stagewise decrease of the number of available conformations may not provide an adequate description of folding kinetics as they actually replace entropic barriers by equally insurmountable energetic ones.

These arguments help us understand better the results of the present work. Compactization itself may not make folding faster because it does not decrease energy barriers on the way to the native conformation. Moreover, it is possible that compactization may considerably increase the barriers. Indeed, compactization decreases the number of conformations available to the chain and, therefore, prohibits some conformations which would form a pathway with lower barriers. The fact that compactization does slow down the folding of real proteins can be deduced from the following observations. In the absence of a compact intermediate the dependence of the folding rate on the concentration of a denaturant is linear. When a compact kinetic intermediate emerges, the dependence becomes nonlinear, and the actual folding rate appears to be smaller than the value corresponding to the linear dependence in the absence of a compact intermediate (Matouschek *et al.*, 1990). These experimental results show directly that compactization slows down folding of proteins.

The problem of effective folding is closely related to the question about the character and magnitude of typical free energy barriers on the way to the native state. As was pointed out before, this includes both entropic and enthalpic components, i.e., the question of the number *and energies* of the relevant conformations. Such transition states were characterized in detail in recent lattice model simulations of 27-monomer chains (Sali *et al.*, 1994b). There is considerable evidence (Abkevich *et al.*, 1994a; Guo & Thirumalai, 1994) that folding kinetics of longer chains involves a nucleation-growth mechanism, with a specific nucleus (Abkevich *et al.*, 1994a) as a low-barrier transition state.

A new paradigm of protein folding, a modern version of the Levinthal paradox, was introduced recently when Wolynes and co-workers (Bryngelson & Wolynes, 1989; Goldstein *et al.*, 1992) pointed out that folding must take place above the so-called glass transition temperature T_c. Under this condition low-energy non-native conformations are not populated and do not serve as traps for folding. It was shown, however (Shakhnovich & Gutin, 1990), that only a vanishingly small fraction of random sequences can have a thermodynamically stable native structure at $T > T_c$. This paradigm is a more consistent formulation of the intrinsic paradox in protein folding and directly implies that protein sequences must have been evolutionary selected to be able to fold, a point made by Levinthal in his original publication (Levinthal, 1969). A possible solution of this paradox was suggested recently (Shakhnovich, 1994; Abkevich *et al.*, 1994a), where it was shown that sequences designed to have their native state as a pronounced energy minimum have kinetic access to this native conformation. This is in line with the original theoretical concepts (Bryngelson & Wolynes,

1989; Goldstein *et al.*, 1992; Shakhnovich & Gutin, 1993a) since such designed nonrandom sequences have stable native structures at temperatures above T_c, resolving therefore the contradiction between thermodynamics and kinetics which is a characteristic feature of random sequences.

ACKNOWLEDGMENT

We are grateful to Oleg Ptitsyn for many important discussions and to Mike Morrisey for editorial help.

REFERENCES

Abkevich, V. I., Gutin, A. M., & Shakhnovich, E. I. (1994a) *Biochemistry 33*, 10026−1036.
Abkevich, V. I.; Gutin, A. M., & Shakhnovich, E. I. (1994b) *J. Chem. Phys. 101*, 6052−6062.
Baldwin, R. L. (1989) *Trends Biochem. Sci. 14*, 291−294.
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.
Camacho, C. J., & Thirumalai, D. (1993a) *Proc. Natl. Acad. Sci. U.S.A. 90*, 6369−6372.
Camacho, C. J., & Thirumalai, D. (1993b) *Phys. Rev. Lett. 71*, 2505−2508.
Chan, H. S., & Dill, K. A. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 6388.
Chan, H. S., & Dill, K. A. (1994) *J. Chem. Phys. 100*, 9238−9257.
Creighton, T. (1992) *Proteins. Structure and Molecular Properties*, W. H. Freeman & Co., New York.
De Dominicis, C., Orland, H., & Lainee, F. (1985) *J. Phys. Lett. 46*, L463.
De Gennes, P. G. (1979) *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca, NY.
De Gennes, P. G. (1985) *J. Phys. Lett. 46*, L639−L642.
Dill, K. (1985) *Biochemistry 24*, 1501−1509.
Dinner, A., Sali, A., Karplus, M., & Shakhnovich, E. (1994) *J. Chem. Phys. 101*, 1444−1451.
Elöve, G. A., Chaffotte, A. F., Roder, H., & Goldberg, M. E. (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.
Fukugita, M., Lancaster, D., & Mitchard, M. G. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 6365−6368.
Gilmanshin, R., & Ptitsyn, O. B. (1987) *FEBS Lett. 223*, 327−329.
Goldberg, M., Semisotnov, G., Friquet, B., Kuwajima, K., Ptitsyn, O. B., & Sugai, S. (1990) *FEBS Lett. 263*, 51−56.
Goldstein, R., Luthey-Schulten, Z. A., & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 9029−9033.
Grosberg, A. Yu., & Shakhnovich, E. I. (1986) *Sov. Phys.−JETP 64*, 1284−1290.
Grosberg, A. Yu., Nechaev, S. K., & Shakhnovich, E. I. (1988) *J. Phys. (Paris) 49*, 2095−2100.
Guo, Z., & Thirumalai, D. (1994) *Biopolymers* (in press).
Harrison, S. C., & Durbin, R. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4028−4030.
Hilhorst, H. J., & Deutch, J. M. (1975) *J. Chem. Phys. 63*, 5153−5161.
Jackson, S. E., & Fersht, A. (1991) *Biochemistry 30*, 10428−10435.
Karplus, K., & Weaver, D. L. (1979) *Biopolymers 18*, 1421−1437.
Karplus, M., & Shakhnovich, E. (1992) in *Protein Folding* (Creighton, T. E., Ed.) Chapter 4, W. H. Freeman and Co., New York.
Khorasanizadeh, S., Peters, I. D., Butt, T. R., & Roder, H. (1993) *Biochemistry 32*, 7054−7063.
Kim, P., & Baldwin, R. (1990) *Annu. Rev. Biochem. 51*, 459.
Kolinski, A., Skolnick, J., & Yaris, R. (1986) *J. Chem. Phys. 85*, 3585−3587.
Kopper, G., & Hilhorst, H. (1987) *Europhys. Lett. 3*, 1213−1217.
Kuwajima, K., Yamaya, H., Miwa, S., Sugai, S., & Nakamura, T. (1987) *FEBS Lett. 221*, 115−118.
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, 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.
Miller, R., Danko, C. A., Fasolka, M. J., Balazs, A. C., Chan, H. S., & Dill, K. A. (1992) *J. Chem. Phys. 96*, 768−780.
Miyazava, S., & Jernigan, R. L. (1985) *Macromolecules 18*, 534−552.
Orland, H., Itzykson, C., & De Dominicis, C. (1985) *J. Phys. Lett. 46*, L353−L355.
Poland, D., & Scheraga, H. A. (1970) *Theory of Helix-Coil Transitions in Biopolymers: Statistical Mechanical Theory of Order-Disorder Transitions in Biological Macromolecules*, Academic Press, New York.
Rey, J., & Skolnick, J. (1991) *Chem. Phys. 158*, 199−203.
Sali, A., Shakhnovich, E. I., & Karplus, M. (1994a) *J. Mol. Biol. 235*, 1614−1636.
Sali, A., Shakhnovich, E. I., & Karplus, M. (1994b) *Nature 369*, 248−251.
Shakhnovich, E. I. (1994) *Phys. Rev. Lett. 72*, 3907−3909.
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) *Europhys. Lett. 9*, 569−574.
Shakhnovich, E. I., & Gutin, A. M. (1990) *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.
Shakhnovich, E., Farztdinov, G., Gutin, A. M., & Karplus, M. (1991) *Phys. Rev. Lett. 67*, 1665−1668.
Skolnick, J., & Kolinski, A. (1991) *J. Mol. Biol. 221*, 499−531.
Socci, N., & Onuchic, J. (1994) *J. Chem. Phys. 101*, 1519−1528.
Sosnick, T., Mayne, L., Hiller, R., & Englander, S. W. (1994) *Nature Struct. Biol. 1*, 149−156.
Stanley, H. E. (1971) *Introduction to Phase Transitions and Critical Phenomena*, Oxford University Press, New York and Oxford.
Weaver, D., & Karplus, M. (1994) *Protein Sci. 3*, 650−668.
Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697−701.

BI941997J