Protein Science (1995), *4*:1167–1177. Cambridge University Press. Printed in the USA.

Domains in folding of model proteins

V.I. ABKEVICH, A.M. GUTIN, AND E.I. SHAKHNOVICH
Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138

(RECEIVED December 22, 1994; ACCEPTED March 27, 1995)

Abstract

By means of Monte Carlo simulation, we investigated the equilibrium between folded and unfolded states of lattice model proteins. The amino acid sequences were designed to have pronounced energy minimum target conformations of different length and shape. For short fully compact (36-mer) proteins, the all-or-none transition from the unfolded state to the native state was observed. This was not always the case for longer proteins. Among 12 designed sequences with the native structure of a fully compact 48-mer, a simple all-or-none transition was observed in only three cases. For the other nine sequences, three states of behavior—the native, denatured, and intermediate states—were found. The contiguous part of the native structure (domain) was conserved in the intermediate state, whereas the remaining part was completely unfolded and structureless. These parts melted separately from each other.

Keywords: lattice models; Monte Carlo simulations; protein domains; protein folding

Theoretical understanding of thermodynamic and kinetic aspects of protein folding has advanced significantly in the last few years (reviewed by Karplus & Shakhnovich, 1992; Chan & Dill, 1993; Frauenfelder & Wolynes, 1994). This became possible due to the extensive use of simplified analytical or numeric lattice models. Even though these models are idealized and miss many details of real proteins, they have important advantages as well. Their key advantage is that they allow one to simulate the complete folding process—from random coil to native conformation—without introducing artificial unphysical biases to the native state (Shakhnovich et al., 1991; Skolnick & Kolinski, 1991; Leopold et al., 1992; Miller et al., 1992; Camacho & Thirumalai, 1993; Abkevich et al., 1994a, 1994b; Shakhnovich, 1994; Socci & Onuchic, 1994). In most of these model simulations, starting from random coil conformations, all runs reach unique native conformation. The numerically exact character of a lattice model and the possibility of eventually addressing any meaningful question make such models a valuable tool for understanding the mechanism of protein folding.

In our previous work, we studied the mechanism of fólding of relatively short proteins. It was found that folding is an all-or-none transition (Abkevich et al., 1994a, 1994b), in accord with theoretical analysis (Goldstein et al., 1992; Shakhnovich & Gutin, 1993) and experimental results (the latter reviewed by Privalov, 1979). Certain basic aspects of the folding mechanism, such as the nature of transition state and intermediates, are understood at the level of a lattice model for relatively short proteins (Abkevich et al., 1994a, 1994b).

At this point, a further step can be made to explore more special features of proteins, especially ones that become pronounced for longer chains. Different parts (domains) of longer proteins unfold and refold like small single-domain proteins, giving rise to complex folding and unfolding curves (Privalov, 1982). There has been considerable controversy in the literature about how to define domains in proteins. For example, in a recent review on multidomain proteins (Garel, 1992), seven different definitions of domains in proteins are quoted. Nevertheless, domains defined by different methods very often coincide. For example, in cases where there are crystallographically distinct separate domains (Janin & Wodak, 1983), they always behave as thermodynamically separate units (Privalov, 1982) that can be separated by limited proteolysis (Novokhatny et al., 1984). However, there are many cases where thermodynamically distinguishable domains, which fold as separate units ("folding domains"; Miranker et al., 1991), cannot be easily identified as structurally distinct species. Moreover, mutations in the same protein can switch it from one-domain to two-domain thermodynamic behavior (Carra et al., 1994).

Recent advances in the theory of protein folding have made it possible to simulate unbiased fast folding of lattice model proteins of considerable length (up to 100 residues; see Shakhnovich, 1994). The important idea, which makes such simulations possible, is to combine design and folding in the "one pair of hands." The key point is to design a sequence that delivers sufficiently low energy to a given structure, so that one can be certain that this "target" structure represents a pronounced

Reprint requests to: E.I. Shakhnovich, Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138; e-mail: eugene@diamond.harvard.edu.

global minimum for this sequence, i.e., it is separated in energy by a large "stability gap" from misfolded conformations. This requirement is sufficient to ensure reliable folding of designed sequences into their native conformations (Goldstein et al., 1992; Shakhnovich & Gutin, 1993; Sali et al., 1994a; Hao & Scheraga, 1994, 1995). The use of the same force-field for design and for folding simulations allows avoidance of the main difficulty related to the uncertainties in the choice of force-field. The errors in force-field can make the native structure higher in energy relative to the misfolded conformations, which will overemphasize the multiple-minima problem.

Designing sequences for the folding simulations makes the particular choice of the force-field not essential, provided that the native state is a pronounced energy minimum for the designed sequence with the chosen force-field. However, not all force-fields can satisfy this condition, e.g., a simple two-letter HP potential (Chan & Dill, 1994) is not specific enough. Many misfolded conformations will have energy equal to or very close to the native conformation even if the sequence and native conformation are so "optimally chosen" (Yue & Dill, 1995) that the number of HH contacts is maximal. Correspondingly, it turns out to be impossible to design sequences with sufficient energy gaps in the HP model (Shakhnovich, 1994). However, the use of 20 types of amino acids does not encounter such difficulty, so it is possible to design long sequences with large stability gaps between folded and misfolded conformations, which ensures folding into unique stable conformations (Abkevich et al., 1994a, 1994b; Hao & Scheraga, 1994; Shakhnovich, 1994).

In simulating folding of longer chains, we also observed multidomain behavior. In this work, we report the results of folding simulations of longer proteins and characterize their folding domains in two cases: when domains can be identified as separate structural units and when there is no such direct structural separation. In the latter case, we found that the domain behavior is sequence specific. We compare experimental data with the results of our MC simulations.

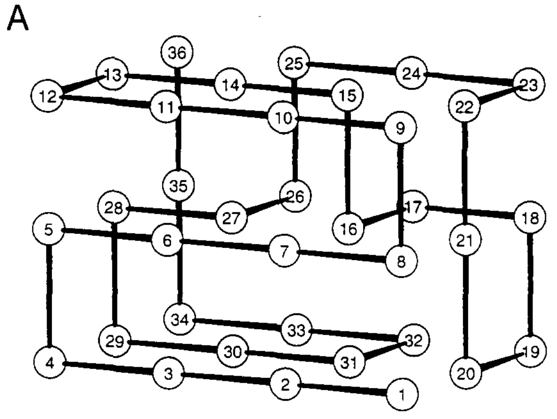
The model

We use a self-avoiding protein chain on an infinite cubic lattice as our basic model. Residues connected by a covalent bond must occupy neighboring lattice sites. We do not consider side chains explicitly. Energy of a conformation is the sum of energies of pairwise contacts between monomers. Two monomers are defined to be in contact if they are neighbors on a lattice and they are not connected by a covalent bond. The energy of a contact depends only on the types of amino acids that are in contact.

The energy parameters for amino acid contact interactions are determined from statistical distributions of contacts in real proteins (Miyazawa & Jernigan, 1985, Table VI). However, protein statistics may provide only relative energies of interactions (Finkelstein et al., 1993). In addition, we also introduced the average contact energy, B_0, which is independent of monomers forming a contact. Thus, the energy of a conformation can be written as follows:

$$E = \sum_{1 \le i < j \le N} [B_0 + B(\xi_i, \xi_j)]\Delta_{ij},$$

where $\Delta_{ij} = 1$ if monomers i and j are lattice neighbors and $\Delta_{ij} = 0$ otherwise, ξ_i defines the type of amino acid residue in the position i, and $B(\xi, \eta)$ is the energy of contact interaction between amino acids of types ξ and η. For real proteins, B_0 might depend on the conditions of folding. However, it should not be too negative, otherwise proteins would aggregate.



B

NKTVVGEPWHCLLFPRRDKNQMSYLTGIAGEDSAAI

Fig. 1. **A:** Randomly chosen maximally compact 36-mer on a cubic lattice. **B:** Native sequence for this conformation.

After choosing a target conformation that serves as a native one in our study, we have to find a sequence that would fold into such a conformation and be stable in it. As was argued previously (Goldstein et al., 1992; Shakhnovich & Gutin, 1993; Sali et al., 1994a, 1994b), this sequence should have the native conformation as a pronounced global energy minimum. To this end, we used the design procedure that is the Monte Carlo (MC) search in sequence space (described in detail in Gutin & Shakhnovich, 1993; Shakhnovich & Gutin, 1993). This procedure generates a number of nonhomologous sequences that have sufficiently low energies in the native conformation, enabling all of them to fold into this conformation. In our simulations we have never found a conformation with an energy lower than the energy of the native conformation, which suggests that the native conformation corresponds to the global energy minimum.

In order to simulate folding, we use the standard MC method (Hilhorst & Deutch, 1975). The move set included corner flips and crankshaft motions and excluded double occupancies of lattice sites. More detailed discussion of the applicability of this algorithm to study folding of lattice model proteins is given in Abkevich et al. (1994a) and Sali et al. (1994a). Different simulation runs begin with different random coil conformations.

Results

The first model protein considered is a 36-mer with the native conformation (Fig. 1) randomly chosen out of more than 84 million fully compact conformations (Pande et al., 1994). There are 40 contacts in the native conformation. To measure the structural similarity between a given conformation and the native state, we used the order parameter—the number of the native contacts in a given conformation.

First, we studied the statistics of the native contacts at equilibrium at different temperatures (Fig. 2). To ensure that equi-

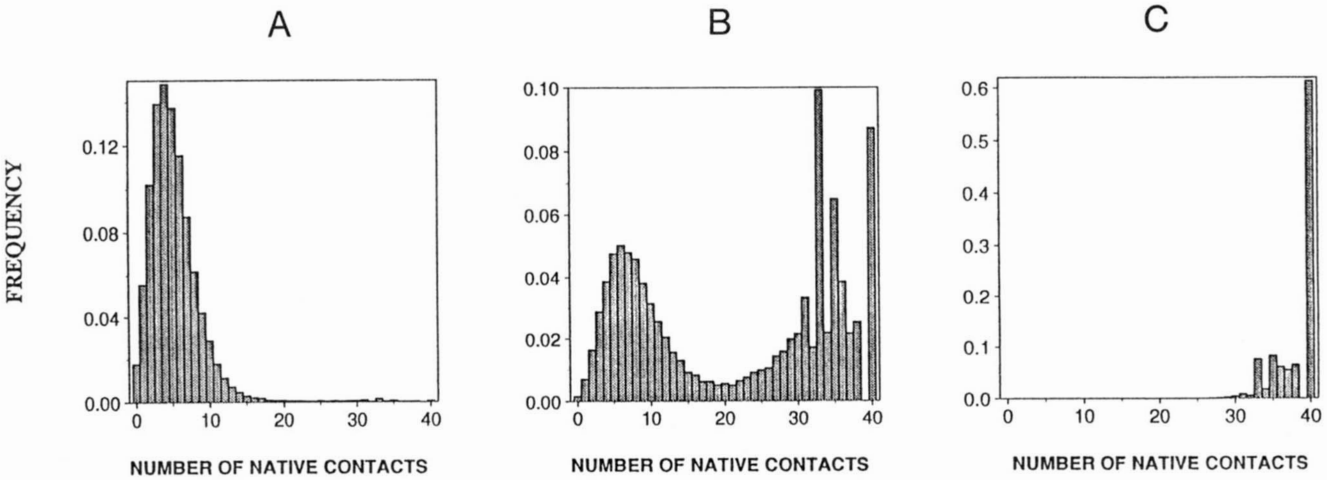


Fig. 2. Distribution of the number of native contacts at equilibrium for the 36-mer shown in Figure 1 at $B_0 = 0.0$ and (**A**) $T = 0.33$, (**B**) $T = 0.28$, and (**C**) $T = 0.21$.

librium is reached, we ran long simulations of 10^8 MC steps at each temperature. The data for histograms were collected every 10^4 steps.[1] At high temperature (Fig. 2A) only one state is dominant. The number of native contacts fluctuates around $n = 5$ at this temperature. This state corresponds to an unfolded protein. At lower temperature, two different states can be distinguished in Figure 2B. The state with the number of native contacts fluctuating around $n = 40$ is, apparently, the native state. The second state is very close to the state that we observe at higher temperature and should also be defined as unfolded. Finally, when temperature is low enough, we again observe in Figure 2C a monomodal distribution with a peak corresponding to the native state. Similar kinds of temperature dependencies for the distributions of native contacts are observed for more than 10 different sequences designed to fold into the same native state (see description of the model). Summarizing the data, we can characterize folding in this case as an all-or-none transition from an unfolded state to the native state. This is in agreement with theoretical views (Bryngelson & Wolynes, 1987; Goldstein et al., 1992; Karplus & Shakhnovich, 1992; Shakhnovich & Gutin, 1993) as well as the wealth of experimental information for small proteins (Privalov, 1979). This is also in line with the nucleation–growth concept of folding (Abkevich et al., 1994a). The nucleation–growth mechanism is typical for all-or-none transitions.

For any chain, the probability for it being in or close to the native conformation can serve as a natural measure of stability of the folded state (Shakhnovich & Gutin, 1990, 1993; Chan & Dill, 1994; Socci & Onuchic, 1994). It can be seen clearly from Figure 2 that conformations with 20 native contacts are least probable, i.e., they correspond to the free energy barrier. Therefore, in this case, when the transition is two-state, we can identify conformations that are "on the native side" of the barrier, i.e., that have more than 20 native contacts as the basin of attraction of the native state. Therefore, we define probability P that our protein at a given temperature is folded as an area on a histogram (Fig. 2) that corresponds to more than 20 native contacts. We use this estimate because it also includes the fluctuation around the native state, not related to complete unfolding. Such fluctuations are always present in proteins at finite temperature in solution, as the results of H–D exchange suggest (see, e.g., Englander & Kallenbach, 1984). The temperature dependence of P is shown in Figure 3. The steepness of the curve in Figure 3 is consistent with the conclusion that the transition between unfolded and folded states is an all-or-none type. Similar temperature dependencies were observed in previous computer simulations (Shakhnovich & Gutin, 1993; Shakhnovich, 1994; Socci & Onuchic, 1994) and in experiments on single-domain proteins (reviewed by Privalov, 1979).

However, deviations from the two-state behavior are often observed in larger multidomain proteins. Such deviations are most clearly pronounced in cases when domains are structurally

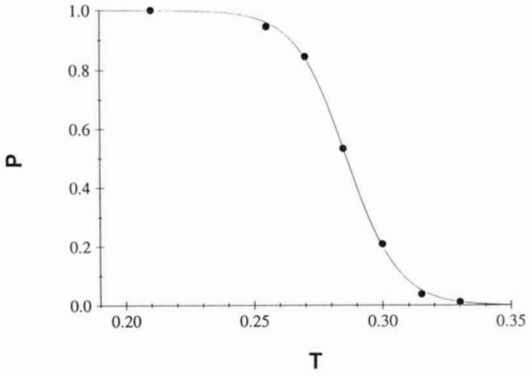
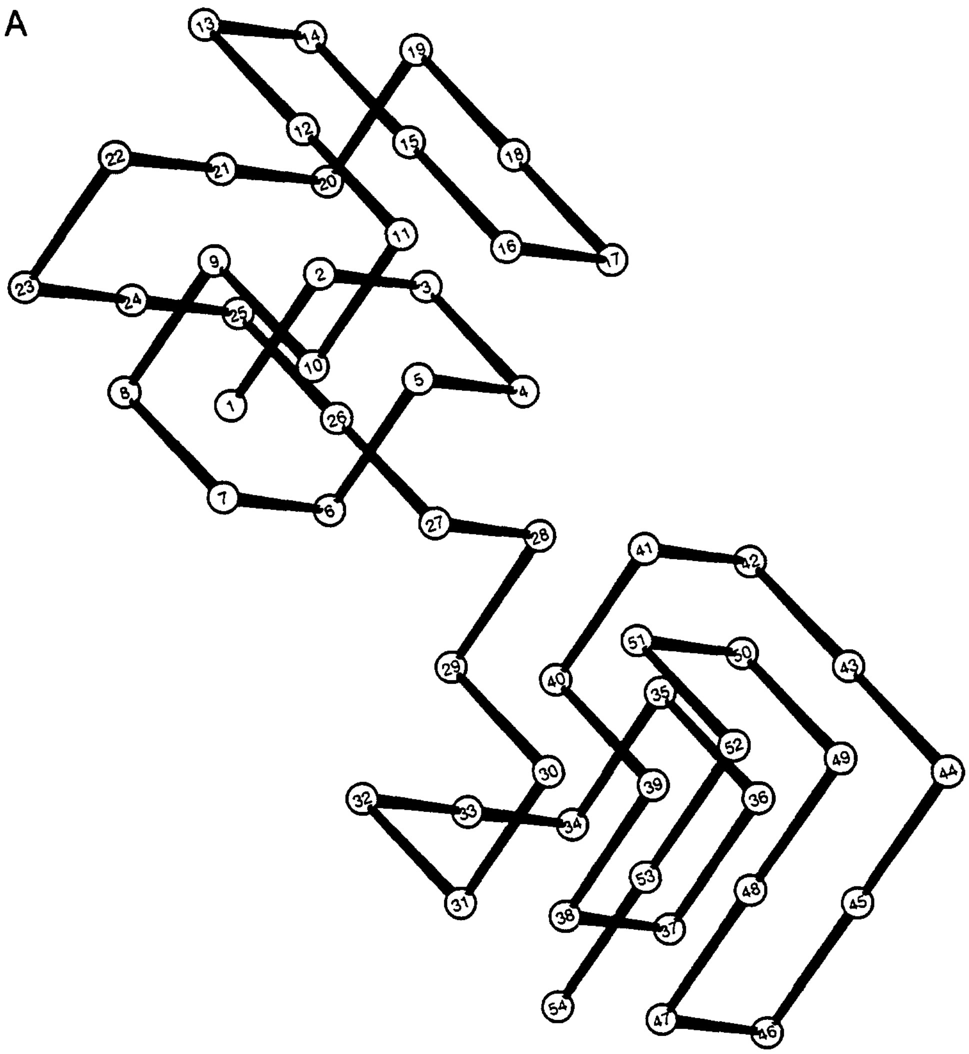


Fig. 3. Temperature (T) dependence of the probability (P) of the 36-mer from Figure 1 being in the folded state at $B_0 = 0.0$.

[1] The reader may have noticed the difference in temperature scales of the present work and previous lattice model simulations from our group (Shakhnovich et al., 1991; Abkevich et al., 1994b; Sali et al., 1994a; Shakhnovich, 1994). This is due to the fact that in previous studies we scaled the parameter set to have $\langle B^2 \rangle = 1$, whereas in this study we use parameters directly as published in Miyazawa and Jernigan (1985, Table VI). The only difference is a constant factor in the temperature scale.

distinct, and each of them undergoes all-or-none transitions independently (Privalov, 1982). A lattice-model example of such a two-domain protein, a 54-mer with two structurally separated structural units (two randomly chosen fully compact 27-mers), is shown in Figure 4. It should be mentioned that this protein model is not fully compact (it has 56 contacts, 8 fewer than the fully compact 54-mer). This lattice protein can be used as a model of real proteins with structurally separated domains. In this case, the presence of domains is clear from simple visual inspection.

The temperature dependence of an equilibrium between the folded and the unfolded states for this 54-mer is quite different than for a 36-mer. First, consider distribution of native contacts at different temperatures (Fig. 5). Comparison of Figure 2 and Figure 5 suggests that at a high temperature (unfolded state dominates) and at a low temperature (native state dominates), histograms for the 54-mer and for the 36-mer are similar. However, at an intermediate temperature, in a simple 36-mer we see an equilibrium between these two states only (see Fig. 2), whereas in Figure 5 the new peak can be clearly distinguished, indicative of the existence of the third state. At the same time, each domain separately undergoes an all-or-none transition, i.e., distributions of the native contacts, if plotted separately for each domain, exhibit bimodal shapes similar to those that are shown in Figure 2. The reason for the appearance of the new peak in Figure 5B is that the folding of two domains of our model protein is not cooperative because interaction between them is weak. The peak in Figure 5B centered around 35 native contacts corresponds to the state when one domain is folded and another one is unfolded. The two domains of the model protein shown in Figure 4 have different thermal stability (as is often the case in real proteins). This becomes clear upon inspection of the melting curves for both domains, shown in Figure 6. Folding temperatures, T_f, are quite different for the two domains. For the first domain (residues 1–27, see Fig. 4) $T_f = 0.18$ and for the second domain (residues 28–54) $T_f = 0.25$.



B

EPQHMKNKQPVPGAQWYPIFQTSKLRDTWQPAPSFENQKMHGKVKYPQLDRIPQ

Fig. 4. **A:** 54-mer with two clearly separated domains. **B:** Native sequence for this conformation.

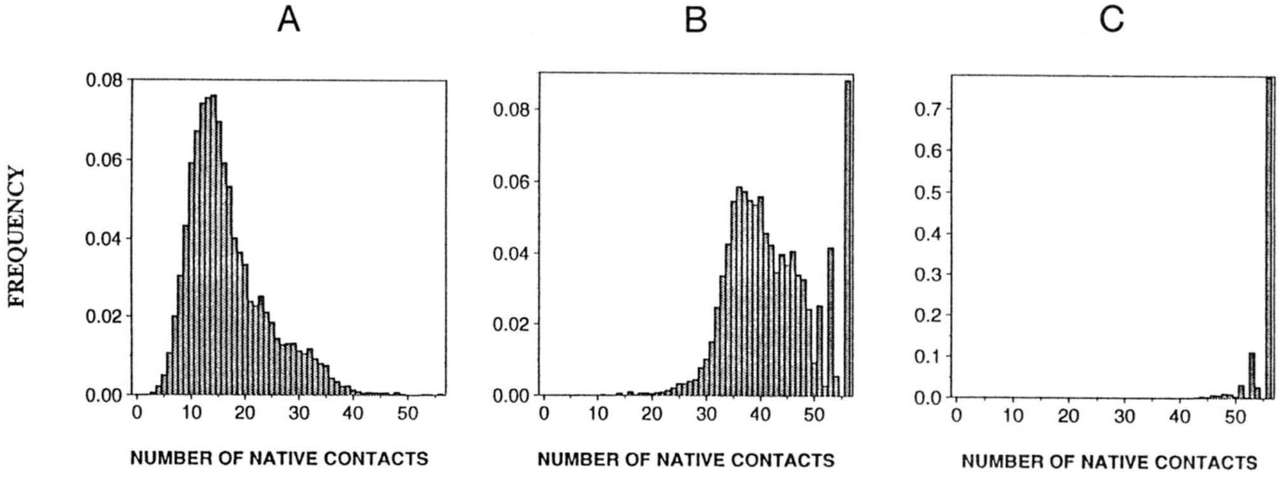


Fig. 5. Distribution of the number of native contacts at equilibrium for the 54-mer from the Figure 4 at $B_0 = 0.0$ and **(A)** $T = 0.3$, **(B)** $T = 0.23$, and **(C)** $T = 0.15$.

Independence of domains in this example can be seen from the thermal stability of a domain being independent of whether another domain is folded or not. We came to this conclusion by studying long (up to 10^8 MC steps) trajectories recorded at $T = 0.24$. All recorded conformations were divided into two groups. To the first group belonged conformations with the second domain folded (i.e., possessing more than 22 native contacts). To the second group belonged conformations with the second domain unfolded (having less than 22 native contacts). Then we studied the statistics of the distribution of native contacts in the *first* domain separately for the ensemble of conformations from the first group and for the ensemble of conformations from the second group. (We call this approach "computational chromatography.") Statistics of the distribution of native contacts in the *first* domain derived from the first group (Fig. 7A) are related to the case when the second domain is folded. Figure 7B illustrates the opposite situation – the statistics of the distribution of the native contacts in the first domain when the second domain is unfolded (collected from the conformations within the second group). The difference between Figure 7A and B is below the noise level. This proves that folding of a model polypeptide with strongly structurally separated domains is not cooperative. This lattice protein can be considered an example of the extreme case, when domains are structurally separated to become totally independent thermodynamically.

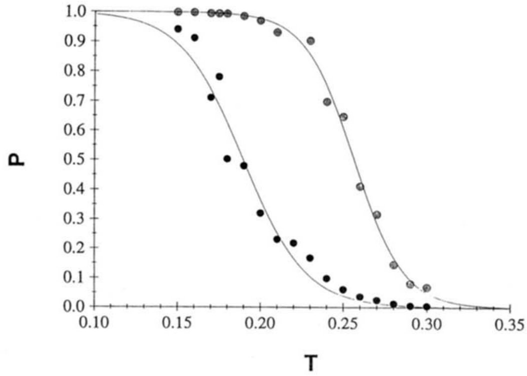


Fig. 6. Temperature (T) dependence of the probability (P) of domains of the 54-mer being in the folded state at $B_0 = 0.0$. Black circles, 27-mer with residues 1–27; gray circles, 27-mer with residues 28–54, as numbered in Figure 4.

The previous case dealt with the model structure where existence of domains was obvious for structural reasons. However, as we mentioned before, domains may exist thermodynamically as separately melting units but may not be readily identifiable as structurally separate objects. Our next lattice model protein is an example of such a situation. A maximally compact 48-mer, randomly chosen out of more than 10^{11} maximally compact conformations (Pande et al., 1994), is shown in Figure 8. Although the two-domain structure of the previous model protein is obvious, this 48-mer looks structurally homogeneous.

There are 57 contacts in the native conformation. Twelve different nonhomologous sequences were designed (Table 1), each of them having the global energy minimum in the conformation shown in Figure 8. The MC design algorithm minimizes energy of a native (target) conformation relative to energies of misfolded conformations, i.e., maximizes the stability gap (Gutin & Shakhnovich, 1993; Shakhnovich & Gutin, 1993). For this reason, it tends to place the most strongly interacting amino acids in the interior, where they can form more contacts. The strongest interactions in Miyazawa–Jernigan parameters from Table VI of Miyazawa and Jernigan (1985) are between charged groups, and therefore in sequences 1–10 in our Table 1, charged groups are buried. This may be adequate to describe the situation in membrane proteins, but in water-soluble proteins, hydrophobic groups are buried and polar ones are exposed. In order to check whether our results are sensitive to this issue, we designed two other sequences by imposing an additional condition penalizing the placement of polar groups inside. As a result, such constrained design generated sequences 11 and 12 in Table 1, which also have low energies in the target state, fold into it, and are stable in this conformation but have hydrophobic groups inside. This is not surprising because hydrophobic interactions are also strong in the parameter set that we used,

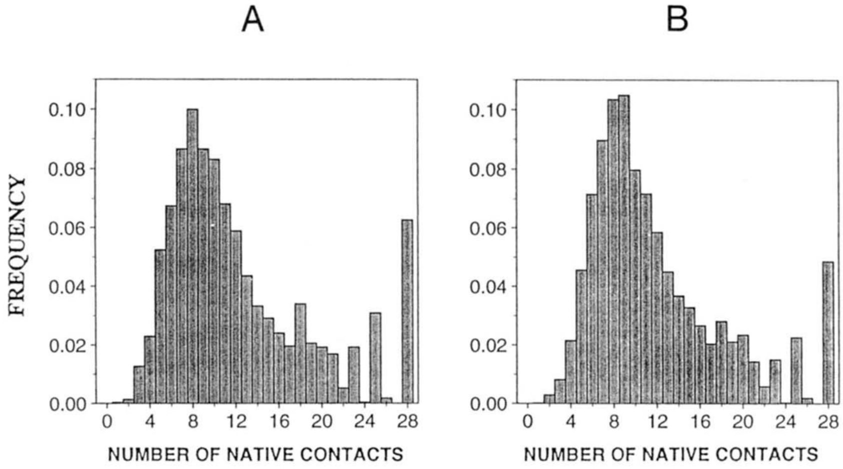


Fig. 7. Distribution of the number of native contacts over all conformations at equilibrium for the 27-mer with residues 1-27 (as numbered in Fig. 4) at $T = 0.24$ and $B_0 = 0.0$, when another 27-mer (residues 28-54) is (**A**) folded or (**B**) unfolded.

so the sequences with buried hydrophobic groups are only marginally less stable, on average, than sequences 1-10 from Table 1. As will be clear below, all of our qualitative conclusions remain the same for sequences with polar groups inside (1-10) and for sequences with hydrophobic groups inside (11, 12).

The probability distributions for the number of native contacts for sequences 1 and 12 from Table 1 at three different temperatures are shown in Figure 9. One can clearly distinguish three peaks for both sequences at an intermediate temperature, which makes this histogram closer to the one shown in Figure 5 (two-domain uncooperative transition) than to the one shown in Figure 2. This suggests that our 48-mer may consist (thermodynamically) of at least two domains. If these domains have different stabilities, only the more stable one will be folded at an intermediate temperature, and therefore contacts constituting this domain will dominate in the intermediate state. If the hypothesis about the two-domain character of folding of our 48-mer is correct, we expect that contacts that are formed in the intermediate state will be spatially localized as well as contacts that are not formed in the intermediate state. In other words, each domain can be identified in this case as a contiguous (in space, not necessarily along the sequence!) substructure formed by contacts

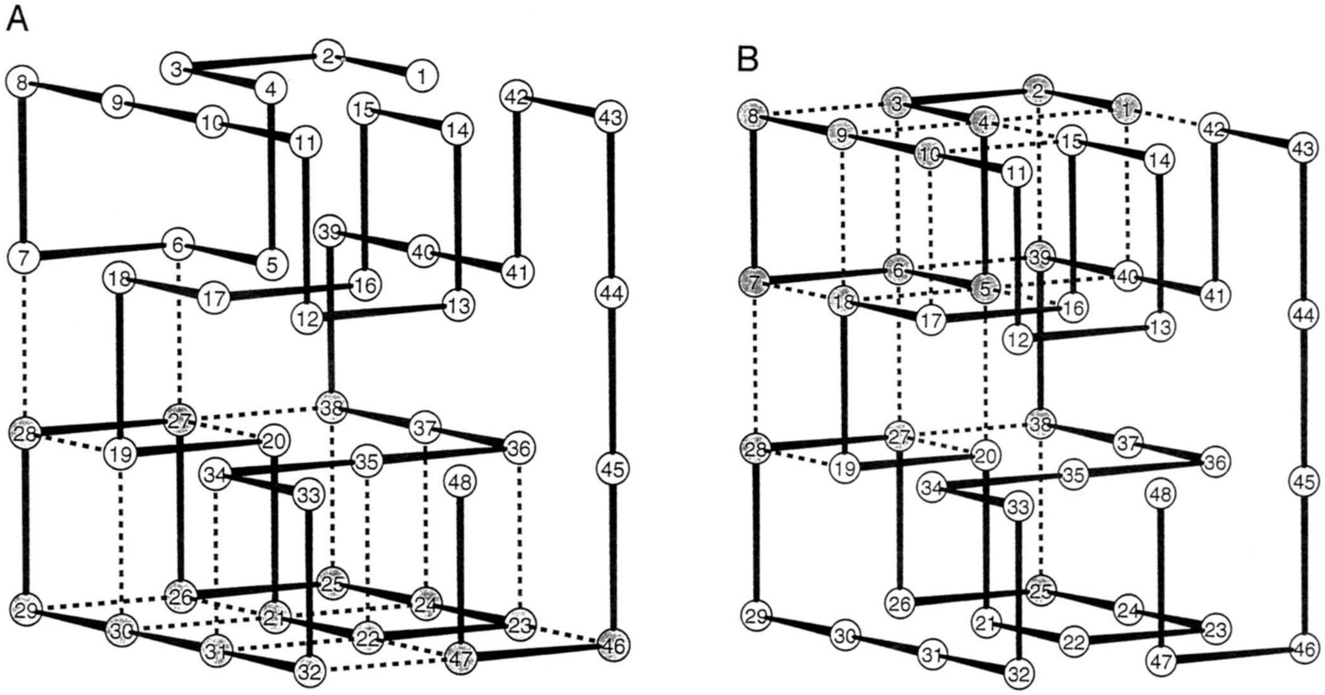


Fig. 8. Randomly chosen maximally compact 48-mer on a cubic lattice. **A:** Division into domains for sequences 1, 2, 3, 5, 6, 9, and 10 from Table 1. Native contacts of the smaller domain are shown by dashed lines. White circles correspond to the larger domain, shaded circles to the smaller one. Circles are shaded if all of their contacts are within the smaller domain. **B:** Division into domains for sequence 4 from Table 1. The same gray scale notation as in A is used to identify domains.

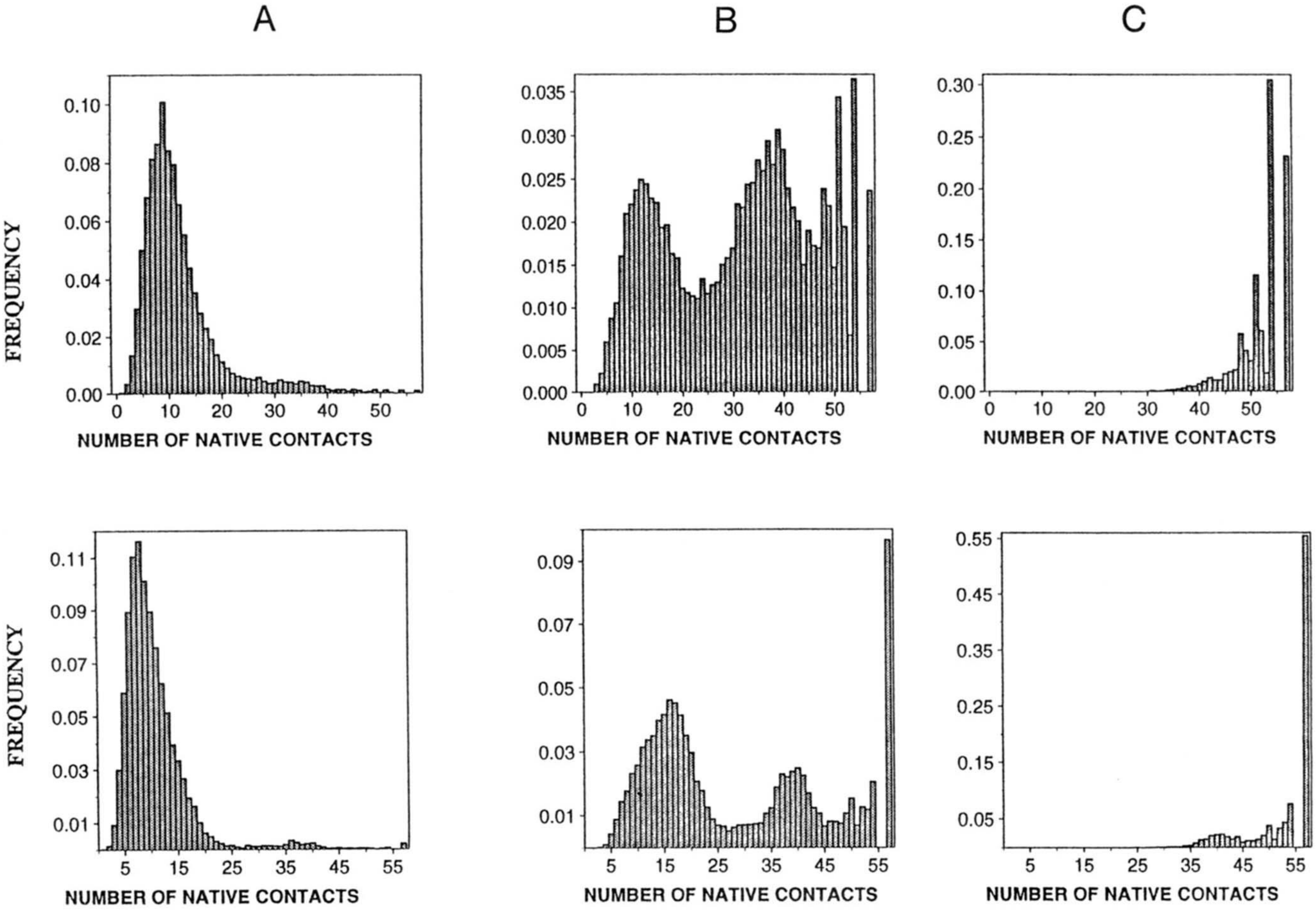


Fig. 9. Distribution of the number of native contacts at equilibrium for the two 48-mer sequences from Table 1 at $B_0 = 0.1$. Upper plots: sequence 1, **(A)** $T = 0.27$, **(B)** $T = 0.24$, and **(C)** $T = 0.20$. Lower plots: sequence 12, **(A)** $T = 0.25$, **(B)** $T = 0.22$, and **(C)** $T = 0.19$.

that are present or absent at the intermediate temperature, $T = 0.24$. We can also expect that both of these groups of contacts will undergo all-or-none transitions during folding.

Therefore, in order to identify folding domains structurally, we should study conformations belonging to the intermediate state at $T = 0.24$, i.e., having between 26 and 44 native contacts.

Table 1. *Nonhomologous sequences designed by the Monte Carlo method*

1	HVIENLTAKQVPEVRKRDGNMSSWLFLGCFGAQTDSPILYEYAKDATR
2	KVLEQIAVSDSGDSKRGTIPFYTGALMAVLHYTFQRNWLNEECKDAPR
3	DVLKGIWVQRYSKNEDTPFTFYSGALAMVLPASQRDGILTRKNEHACE
4	FVLLSIAVIMVHKLFDYAQTDPSTARATERQNYPRDGWLGKGCEKSNE
5	VKDLPKWSFLAAKAIEVMLQGTHSYGNVGFASTIRDPNEYRLTERCQD
6	KVLQNIVVPQYGEPRKFYINGSTTAFMLCAAATLDRHWLEEDSKDGSR
7	LKDLQRATILVCKVIEFAGPHTSWYYQAPTGNSGRDMREFKLVESAND
8	RVLDQIAVSDSADSKRTGWFFYNGALLAIMHYTPQKGVLTEECKENPR
9	HVCEPLFARQSNESKRYQMTLTTFLILWAIGANPDRGVVYDESKDAGK
10	DLVRGLVLQHPNKSEDYFWGFQTCAIALAMYANPRDGIVTRKSEKSTE
11	RAIEFLTASPSWGSRFMGPMPISYVRIRERLIMWFGPVIQGESGGSLF
12	KAIEMLGAKTGGESKWGTGWIPYLLVLGVAGMTGMKMIIWEESKEPWK

To this end we again used "computational chromatography," generating long MC trajectories and selecting conformations of the intermediate state, as defined above. We determined the probability for each native contact to be found in the intermediate state (see Fig. 10). One can see in Figure 10 that certain contacts exist with high probability in the intermediate state, whereas other contacts are only marginally present in the intermediate state. Contacts that occurred with high frequency (above the dashed line) were attributed to the first domain; contacts having low frequency (below the dashed line) were attributed to the second domain. Identification of high- and low-frequency contacts in the native conformation allowed us to determine domains for different sequences (see below). Most importantly, the residues belonging to each domain as determined from the diagram in Figure 10 constitute spatially contiguous substructures. This makes the definition of domains meaningful and intuitive.

The resulting division of the 48-mer into two domains for sequence 1 (Table 1) is shown in Figure 8A. One of these domains consists of 39 native contacts, another of 18 native contacts. The main part of the smaller domain is comprised of a single loop (residues 21–32, as they are numbered in Fig. 8). The reader should understand that, unlike the case of the 54-mer, which we discussed above, such division can be only approximate because the domains are not so clearly distinct in this case. Such is true

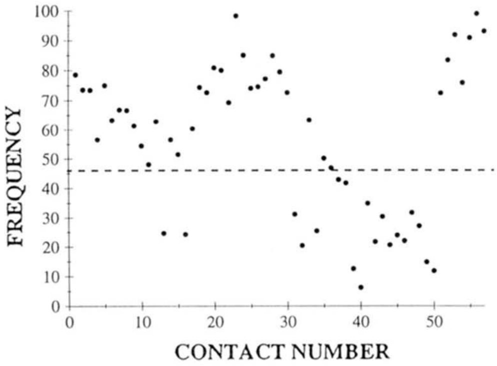


Fig. 10. Frequencies with which individual native contacts occur in the intermediate (having between 26 and 44 native contacts) in a long (10^8) MC trajectory at $T = 0.24$ at $B_0 = 0.1$. Each of the 57 native contacts was ascribed a number for identification purposes.

even in real proteins, where the division of a structure into domains is sometimes a subjective process that is done in different ways by different authors.

If the division into domains shown in Figure 8A is correct, then each domain will melt as a cooperative unit. For each domain, the distribution of the number of the native contacts at different temperatures should be similar to those shown in Figure 2 and typical for an all-or-none transition. For example, this dependence on the larger domain is shown in Figure 11. This implies that the third peak in Figure 9B is due to the same reason as the third peak in Figure 5B, namely the two-domain structure. For each domain, the measurements of thermal stability were made. The melting curves for both domains are shown in Figure 12. We observe two S-shaped curves typical of all-or-none transitions. For the larger domain, $T_f = 0.25$, and for the smaller one, $T_f = 0.23$. Not surprisingly, folding temperatures for intersecting domains turned out to be closer to each other than for separate ones. We should mention that, according to the simple analytical estimate, the width of the folding transition in Figures 3, 6, and 11 should be inversely proportional to the length of the cooperative unit undergoing an all-or-none transition as a whole (Lifshitz et al., 1978). In the case of the 54-mer, domains are of the same length. Accordingly, widths of folding transition are approximately the same for both domains. For the 48-mer, however, one domain is twice as large as the other one. As a result, the width of the folding transition for the smaller domain is two times larger than for the larger domain.

The two-domain structure implies that folding of a protein is not an all-or-none process. However, interactions between domains lead to a certain degree of cooperativity of folding (Shakhnovich & Finkelstein, 1989). Apparently, for our 48-mer, we cannot expect complete independence of domains—a characteristic of the 54-mer discussed earlier. In order to evaluate the degree of dependence between different domains for the 48-mer model, we studied the probability distribution for the number of native contacts in the shorter domain in the cases when the larger one is formed and in the case when it is unfolded. This is done using the same method of computational chromatography as described above for the 54-mer model. We studied long equilibrium MC trajectories taken at $T = 0.24$ and divided all recorded conformations into two classes: conformations with the first (larger) domain formed, and conformations when this domain was unfolded. In Figure 13, the distributions of the number of native contacts for the smaller domain, when the large one is folded (Fig. 13A) and unfolded (Fig. 13B), are compared.

The difference between Figure 13A and B is clear. This suggests that folding of the smaller domain is possible only if the larger one is folded. This 48-mer can be viewed as a model of a protein with strongly interacting domains, as, for example, staphylococcal nuclease (Carra et al., 1994) or human plasminogen (Novokhatny et al., 1984).

We studied folding of model proteins with sequences shown in Table 1 at different temperatures. All of the sequences fold into the structure shown in Figure 8 and, under certain temperature conditions, are stable in this conformation.

Folding of 9 of 12 sequences exhibits similar types of temperature dependence for the distribution of native contacts, as shown in Figure 9, which means that they also have a multi-

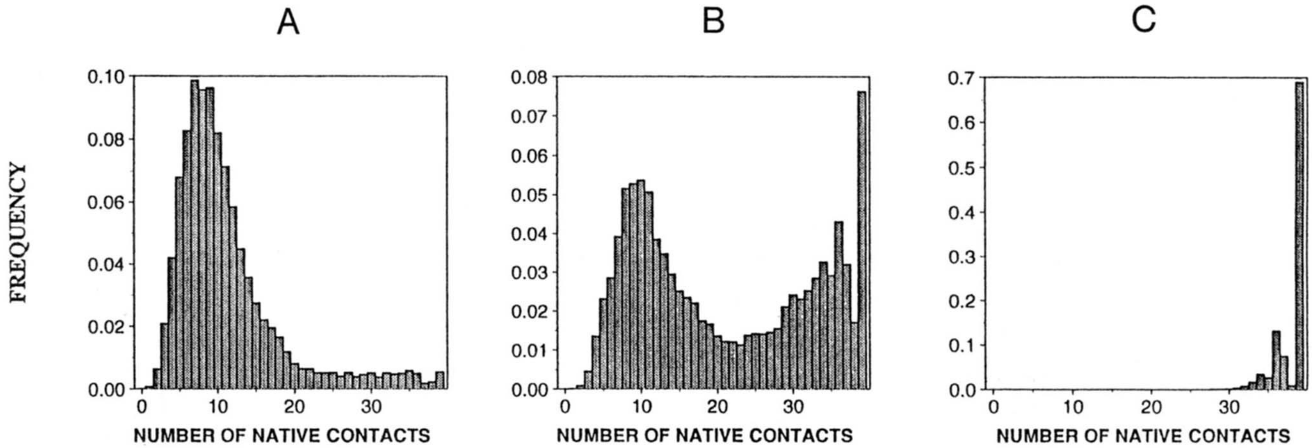


Fig. 11. Distribution of the number of native contacts at equilibrium for the first sequence from Table 1 at $B_0 = 0.1$ and (**A**) $T = 0.27$, (**B**) $T = 0.26$, and (**C**) $T = 0.2$ for the larger domain in Figure 8A.

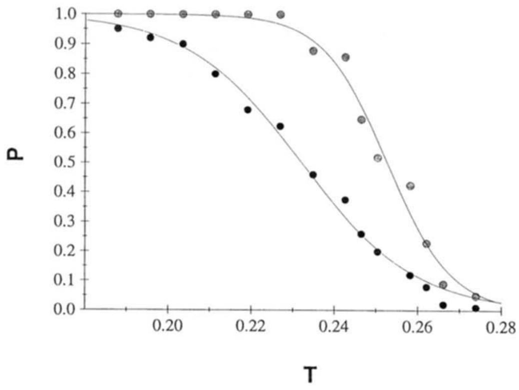


Fig. 12. Temperature (T) dependence of the probability (P) of being in the folded state for different domains of a 48-mer (see Fig. 8A). Black circles, smaller domain; gray circles, larger domain for the first sequence from Table 1. $B_0 = 0.1$.

domain folding character. However, sequences from Table 1 (7, 8, and 11) behave differently, with only two peaks present on their temperature dependence for the distribution of native contacts (shown for sequence 8 in Fig. 14A and for sequence 11 in Fig. 14B), which suggests that they have a single-domain cooperative folding transition.

For eight sequences from Table 1 we found the division into domains to be the same as for the first sequence (shown in Fig. 8A). However, for the fourth sequence from Table 1, which also has a multidomain character of folding, division into domains is different, as shown in Figure 8B. This implies that sometimes domains are not solely determined by protein structure but also depend on the sequences that have this structure as the native state. We can see in Figure 8B that the smaller domain consists mainly of a single segment of a polypeptide chain (residues 1–10). Domains should probably consist of contiguous segments of the chain. Further work in this direction is necessary to evaluate the generality of this conclusion.

Discussion

Our results suggest that a simple lattice model can be further used for the description of more complicated aspects of folding such as multidomain behavior, both for proteins with structurally separated domains and for proteins with strongly intersecting domains. As opposed to the case for single-domain proteins, the folding transition in multidomain lattice proteins is not of an all-or-none character, as in real proteins. In addition to the unfolded and native state, the third, intermediate, state exists, in which one domain is folded and another one is unfolded. Different domains may have different folding temperatures, which give rise to complex melting curves such as the ones shown in Figures 6 and 12.

The important feature found in simulations with 48-monomer model proteins is that, for the same native structure, the character of separation into domains and even their existence may depend on the amino acid sequence. There is a wealth of experimental information supporting this conclusion. For example, it was shown for staphylococcal nuclease that a number of mutations can make the protein thermodynamically two-domain, with differences in transition temperatures of up to 20 °C, whereas the wild-type protein behaves like a single-domain protein, having one calorimetric melting peak (Carra et al., 1994). Another studied example is lysozyme, where domains can hardly be identified as structural ones, but folding kinetics studies revealed distinct separate "folding domains" (α-domain and β-domain) having clear differences in folding rate. Moreover, it was determined that lysozymes from different sources (hen-egg [HEL] and horse) that are highly homologous behave quite differently—although HEL behaves like a two-domain protein in thermodynamic melting experiments, horse lysozyme can be well characterized thermodynamically as a single-domain protein (P.L. Privalov, pers. comm.). This is in clear correspondence with our results, where 8 (1, 2, 3, 5, 6, 9, 10, 12) of 12 sequences (Table 1) designed to fold into the native structure shown in Figure 8 have domains as shown in Figure 8A, sequence 4 has the domain structure shown in Figure 8B, and sequences 7, 8, and 11 have no thermodynamically revealed domains at all!

A question may arise as to what extent this conclusion is sensitive to the design procedure used and the sequence of the motifs generated. To address this question we note that sequences 1–10 were generated by an unconstrained design procedure and have polar groups inside. Sequences 11 and 12 were designed to have hydrophobic groups inside and polar groups outside to reproduce the corresponding property of water-soluble proteins. In both cases we observe sequence-dependent domain behavior:

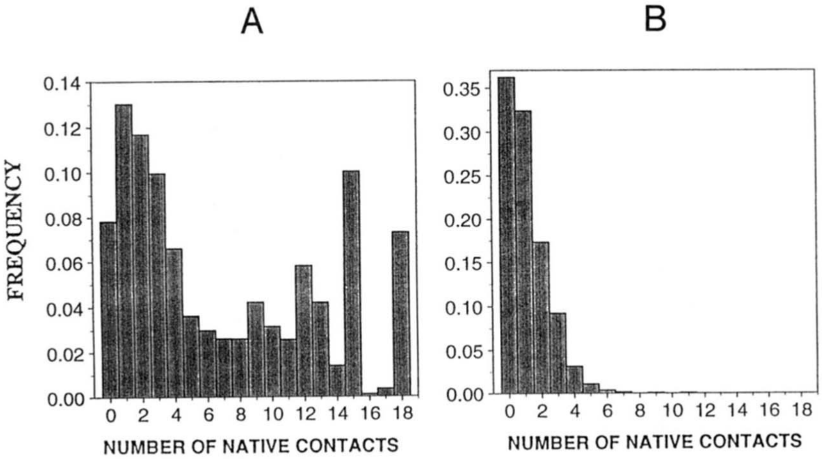


Fig. 13. Distribution of the number of native contacts at equilibrium for the smaller domain in Figure 8A for the first sequence from Table 1 at $T = 0.24$ and $B_0 = 0.1$, when the larger domain is (**A**) folded or (**B**) unfolded. We consider the larger domain to be in the folded state if its conformation has more than 23 native contacts.

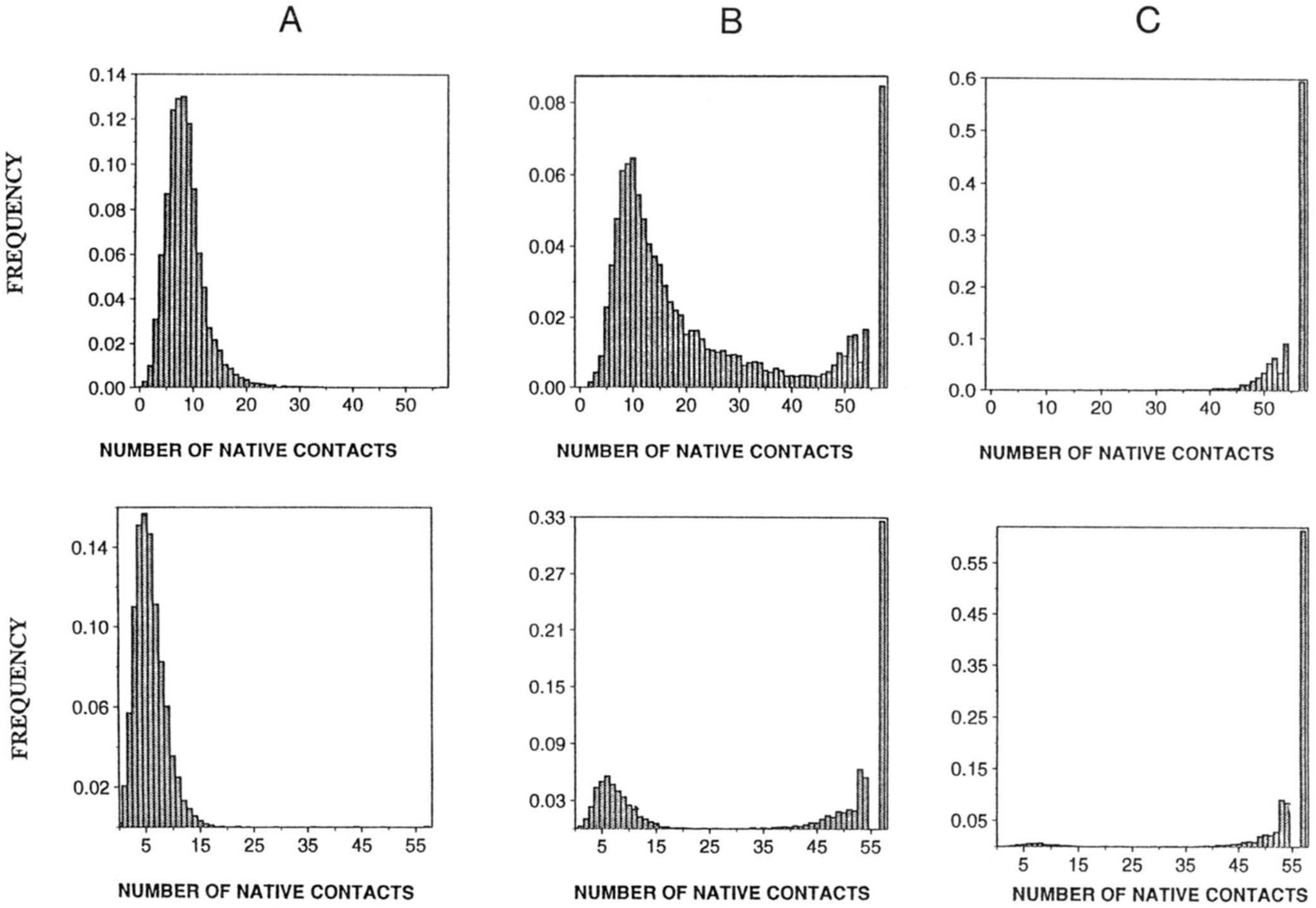


Fig. 14. Distribution of the number of native contacts over all conformations at equilibrium for sequences that do not have multiple domains. $B_0 = 0.1$. Upper plots: sequence 8, (**A**) $T = 0.28$, (**B**) $T = 0.25$, and (**C**) $T = 0.22$. Lower plots: sequence 11, (**A**) $T = 0.25$, (**B**) $T = 0.225$, and (**C**) $T = 0.21$.

sequence 11 does not have domains in its native structure (see Fig. 13B), whereas sequence 12 does (see Fig. 9B). This suggests that the conclusion that domain behavior may be sequence specific is very robust and holds for both patterns of design.

An important question is what features of sequences determine its single- or multidomain behavior? Qualitatively one can suggest that this may be related to the heterogeneity of possible contact energies in the native structure: in the interface between domains, contacts may be less favorable. Also, one may expect that domain structure will still be dependent on the native conformation: conformations where a large contiguous part of the chain is (partly) separated from the remaining chain are more likely to exhibit a multidomain behavior. It remains to be answered how often such conformations can be found in the lattice model proteins and the real proteins. An important goal of theoretical approaches in this area is to design sequences of lattice model proteins that have desired one- or multidomain behavior.

In this work we discussed the three-state thermodynamic behavior of model proteins related to their possible multidomain structure. The three-state thermodynamics was also observed in a number of proteins and studied theoretically as evidence of the molten globule (MG) state (Shakhnovich & Finkelstein, 1982, 1989; Ptitsyn, 1987, 1992). One can get the superficial impression that findings of this work are related to the MG state. Many results of experimental studies suggest that the MG state is a global feature of a protein molecule, i.e., the whole chain becomes non-native when the transition to MG occurs. This can be judged by dramatic global changes in a number of physical characteristics of a molecule, e.g., a coil-like spectrum at 290 nm CD, qualitatively different character of the NMR spectrum, etc. (see, e.g., the review by Ptitsyn, 1992). This is consistent with the concept that the MG state does not possess tight packing of side chains in the core, though it may possess some elements of native-like (albeit, strongly fluctuating) backbone organization. An important experimental signature of the MG state is the pronounced ANS binding (Semisotnov et al., 1991), which is evidence for the existence of loose clusters of hydrophobic groups.

Multidomain behavior differs from that of MG. In the intermediate state of a multidomain protein, part of the molecule is native and another part is unfolded. Evidence for that is that in the partly folded state staphylococcal nuclease does not bind ANS (Carra et al., 1994), whereas the A-state of the same protein binds ANS much more strongly than both the unfolded and the native states.

An interesting question is the relation of domain structure of a chain to its folding kinetics. Folding domains were found to play an important role in the kinetics of folding in lysozyme (Miranker et al., 1991; Radford et al., 1992). In fact, one could suggest that the existence of domains may slow down folding due to the fact that the less stable part of the structure would fold at a much lower temperature, at which folding of the more stable domain will dramatically slow down. This is the subject of further study.

Acknowledgments

This work was supported by the Packard Foundation. We thank Ruvim Breydo and Michael Morrisey for editorial help. The graphic program ASGL by Andrej Sali was used to generate some of the plots.

References

Abkevich VI, Gutin AM, Shakhnovich EI. 1994a. Specific nucleus as a transition state for protein folding: An evidence from lattice model. *Biochemistry 33*:10026-10036.

Abkevich VI, Gutin AM, Shakhnovich EI. 1994b. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *J Chem Phys 101*:6052-6062.

Bryngelson JD, Wolynes PG. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA 84*:7524-7528.

Camacho C, Thirumalai D. 1993. Kinetics and thermodynamics of folding in model proteins. *Proc Natl Acad Sci USA 90*:6369-6372.

Carra JH, Anderson EA, Privalov PL. 1994. Three-state thermodynamic analysis of the denaturation of staphylococcal nuclease mutants. *Biochemistry 33*:10842-10850.

Chan HS, Dill KA. 1993. The protein folding problem. *Phys Today 46*(2): 24-32.

Chan HS, Dill KA. 1994. Transition states and folding dynamics of proteins and heteropolymers. *J Chem Phys 100*:9238-9257.

De Gennes PG. 1979. *Scaling concepts in polymer physics*. Ithaca, New York: Cornell University Press.

Englander SW, Kallenbach NR. 1984. Hydrogen exchange and structural dynamics of proteins and heteropolymers. *Q Rev Biophys 16*:521-655.

Finkelstein AV, Gutin AM, Badretdinov AYa. 1993. Why are the same protein folds used to perform different functions? *FEBS Lett 325*:23-28.

Frauenfelder H, Wolynes PG. 1994. Biomolecules: Where the physics of complexity and simplicity meet. *Physics Today 47*(2):58-64.

Garel JR. 1992. Folding of large proteins: Multidomain and multisubunit proteins. In: Creighton TE, ed. *Protein folding*. New York: W.H. Freeman and Company. pp 127-195.

Goldstein R, Luthey-Schulten ZA, Wolynes PG. 1992. Optimal protein-folding codes from spin-glass theory. *Proc Natl Acad Sci USA 89*:4918-4922.

Gutin AM, Shakhnovich EI. 1993. Ground state of random copolymers and the discrete random energy model. *J Chem Phys 98*:8174-8177.

Hao MH, Scheraga H. 1994. Statistical thermodynamics of protein folding: Sequence dependence. *J Phys Chem 98*:9882-9886.

Hao MH, Scheraga H. 1995. Statistical thermodynamics of protein folding: Comparison of a mean-field theory with Monte-Carlo simulations. *J Chem Phys 102*:1334-1339.

Hilhorst HJ, Deutch JM. 1975. Analysis of Monte-Carlo results on the kinetics of lattice polymer chains with excluded volume. *J Chem Phys 63*:5153-5161.

Janin J, Wodak S. 1983. Structural domains in proteins and their role in the dynamics of protein function. *Prog Biophys Mol Biol 42*:21-78.

Karplus M, Shakhnovich EI. 1992. Protein folding: Theoretical studies of thermodynamics and dynamics. In: Creighton TE, ed. *Protein folding*. New York: W.H. Freeman and Company. pp 127-195.

Miranker A, Radford S, Karplus M, Dobson C. 1991. Demonstration by NMR of folding domains in lysozyme. *Nature 349*:633-636.

Miyazawa S, Jernigan R. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534-552.

Novokhatny VV, Kudinov SA, Privalov PL. 1984. Domains in human plasminogen. *J Mol Biol 179*:215-232.

Pande VS, Joerg C, Grosberg AYu, Tanaka T. 1994. Enumeration of the Hamiltonian walks on a cubic lattice. *J Phys A 27*:6231- 6236.

Privalov PL. 1979. Stability of proteins. *Adv Protein Chem 33*:167-241.

Privalov PL. 1982. Stability of proteins: Proteins which do not present a single cooperative system. *Adv Protein Chem 35*:1-104.

Ptitsyn OB. 1992. The molten globule state. In: Creighton TE, ed. *Protein folding*. New York: W.H. Freeman and Company. pp 243-300.

Radford S, Dobson C, Evans P. 1992. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature 358*:302-307.

Sali A, Shakhnovich EI, Karplus M. 1994a. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J Mol Biol 235*:1614-1636.

Sali A, Shakhnovich EI, Karplus M. 1994b. How does a protein fold? *Nature 369*:248-251.

Semisotnov GV, Rodionova NA, Razgulyaev OV, Uversky VN, Gripas AF, Gilmanshin RI. 1991. Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe. *Biopolymers 31*: 119-128.

Shakhnovich EI. 1994. Proteins with selected sequences fold to unique native conformation. *Phys Rev Lett 72*:3907-3909.

Shakhnovich EI, Farztdinov GM, Gutin AM, Karplus M. 1991. Protein folding bottlenecks: A lattice Monte-Carlo simulation. *Phys Rev Lett 67*: 1665-1668.

Shakhnovich EI, Finkelstein AV. 1982. On the theory of cooperative transitions in proteins. *Dokl Acad Nauk SSSR 243*:1247-1251.

Shakhnovich EI, Finkelstein AV. 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers 28*:1667-1681.

Shakhnovich EI, Gutin AM. 1990. Enumeration of all compact conformations of heteropolymers with quenched disordered sequence of links. *J Chem Phys 93*:5967-5971.

Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA 90*:7195-7199.

Skolnick J, Kolinski A. 1991. Dynamic Monte-Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol 221*:499-531.

Socci ND, Onuchic JN. 1994. Folding kinetics of protein-like heteropolymers. *J Chem Phys 101*:1519-1528.

Weaver D, Karplus M. 1994. Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Sci 3*:650-668.

Yue K, Dill KA. 1995. Forces of tertiary structural organization in globular proteins. *Proc Natl Acad Sci USA 92*:146-150.