Protein Folding Bottlenecks: A Lattice Monte Carlo Simulation

E. Shakhnovich, G. Farztdinov,[a] A. M. Gutin,[a] and M. Karplus
Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138
(Received 28 June 1991)

Results of Monte Carlo simulations of folding of a model "protein," which is a freely joined 27-monomer chain on a simple cubic lattice with nearest-neighbor interactions, are reported. All compact self-avoiding conformations of this chain have been enumerated, and the conformation ("native") corresponding to the global minimum of energy is known for each sequence. Only one out of thirty sequences folds and finds the global minimum. For this sequence, the folding process has a two-stage character, with a rapid noncooperative compactization followed by a slower transition over a free-energy barrier to the global minimum. The evolutionary implications of the results are discussed.

PACS numbers: 87.15.−v, 02.50.+s, 05.50.+q

The determination of the mechanism by which proteins fold to their native three-dimensional structure is one of the most challenging unsolved problems of molecular biophysics. It is not clear whether the native structure is determined thermodynamically or kinetically (i.e., whether the native structure of a protein is the global energy minimum), although the former appears more likely from the wide range of conditions under which the same native structure is obtained. Levinthal [1] argued that the astronomically large number of conformations available to a polypeptide chain make it impossible to find the global minimum by a random search in the conformation space. Various possibilities have been considered to simplify the folding process. In one model [2] the restriction arises from the role of secondary structural elements in determining folding pathways. Alternatively, certain lattice [3] or more detailed model [4,5] simulations of protein folding have introduced a bias in the allowed conformations (e.g., higher turn probabilities in the positions found to have turns in the native state) which guarantee folding to the native structure. In cases where such bias was not introduced, simulations [6,7] often ended up with different structures at different runs.

In this paper we take a different approach. We consider a protein model in which the native conformations are determined by nonlocal interactions without assuming any intrinsic propensities for local order. A 27-monomer chain on a simple cubic lattice with nearest-neighbor interactions is used [8]. The conformational space of this chain includes all extended conformations, as well as the subset of fully compact self-avoiding conformations (CSA, Fig. 1); the latter have been fully enumerated and analyzed for certain sets of interactions numerically [8] and analytically [9]. A range of sequences is examined and Monte Carlo (MC) simulations are used to determine which, if any, sequences are able to fold. The sequence parameters are chosen so that in every case a single CSA conformation is thermodynamically dominant below a transition temperature. A key point of this approach is that for each sequence the conformation of lowest energy is known so that we may determine whether the folding process reaches the global minimum.

The model is a 27-monomer freely jointed chain with unit bond length on a simple cubic lattice. Monomers are positioned on lattice sites. Each configuration can be described by a set of positions of the monomers, r_i. In the simulation, which consists of a random walk with unit steps, all possible conformations, including those which are noncompact or have more than one monomer in a site, are allowed. The latter takes account of the inherently greater flexibility of a real protein chain and avoids topological constraints which significantly slow down the kinetics of folding [10].

The energy of a chain with N monomers has the form

$$E = B_0 \sum_{ij}^{N} \Delta(r_i - r_j) + \sum_{ij}^{N} B_{ij} \Delta(r_i - r_j) + D_2 \sum_{ij}^{N} \delta(r_i - r_j) + D_3 \sum_{ijk}^{N} \delta(r_i - r_j)\delta(r_i - r_k) . \quad (1)$$

Here the nearest-neighbor Kronecker delta Δ equals 1 if $r_j - r_i = 1$ and is 0 otherwise; the site Kronecker delta δ equals 1 if $r_i - r_j = 0$ and is 0 otherwise. The first term in Eq. (1) represents a mean attraction ($B_0 < 0$) between monomers occupying neighboring sites that leads to compactization of the chain [11]. The second term is "sequence specific." We generate a given sequence by selecting the B_{ij} from a random distribution with zero

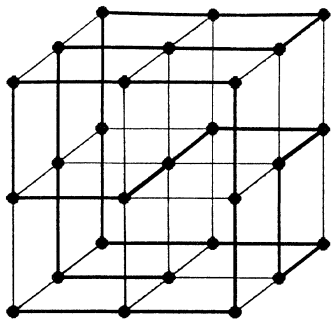

FIG. 1. An example of CSA on a 3×3×3 fragment of a cubic lattice (fat line).

© 1991 The American Physical Society

mean; i.e.,

$$P(B_{ij}) = (2pB^2)^{-1/2}\exp(-B_{ij}^2/2B^2), \quad (2)$$

where B is the standard variance which determines the heterogeneity of the sequence. As mentioned in the introduction, the thermodynamic behavior of heteropolymers with such interactions, which correspond to the "random" energy model, can be determined [8,9]. The last two terms in Eq. (1) introduce energetic penalties for conformations in which two and three links occupy the same site. Thus, although such conformations are allowed in the kinetics of folding, the values of D_2 and D_3 lead to very low thermodynamic probabilities and the ground state is a self-avoiding configuration. The dynamics of a chain was simulated by the standard Metropolis algorithm [12], using the kinetic scheme of Verdier [13]. A monomer with position r_i on the lattice is picked at random and is moved to another position r_i' that is allowed by the polymeric bonds; i.e., $r_i' = r_{i-1} + r_{i+1} - r_i$. Acceptance of the new structure is determined by using the energy expression [Eq. (1)] in the Metropolis algorithm [12]. Overall, about 1% of the moves were accepted in the compact conformations and about 70% were accepted in coil conformation.

The folding of thirty sequences, characterized by sets of B_{ij} generated with the distribution in Eq. (2), was examined. Each run for a given sequence started from an initial conformation that corresponded to an extended (coil) state with random numbers specifying the position of each unit on the lattice. From some preliminary studies, the parameters were taken to be $B_0 = -2$, $B = 1$, $D_2 = 10$, and $D_3 = 14$; all quantities are in units of $k_B T$ with $k_B = 1$. These values yield a unique CSA ground state below the transition temperature with the Boltzmann probability of this state very close to 1. The temperature T in most runs was set equal to unity; however, for certain cases a range of temperatures ($T = 1$–5) was investigated.

Of the thirty sequences which were examined, only three were able to fold to the conformation of the known global minimum and only one exhibited stable and rapid folding to this state, independent of the initial coil configuration. The one "folding" sequence required between 5×10^5 to 5×10^7 MC steps to find the global minimum; test runs extending up to 5×10^8 steps were unsuccessful for the other sequences. The energy of the global minimum of the folding sequence is $E_0 = -82.05$; the other sequences had E_0 values in the range -83.72 to -74.61. We describe first the analysis of the folding sequence and then examine how its potential surface differs from that of the other sequences.

For the folding sequence, fifty simulations were made and all reached the global minimum. To characterize the folding process as a function of the number of Monte Carlo steps, we consider the energy E, the number of nonlocal contacts C (i.e., the contacts other than those with sequence neighbors), and the overlap Q with the lowest-energy conformation; the quantity Q is defined as

$$Q = N_{\text{common}}/N_{\text{total}}, \quad (3)$$

where N_{common} is the number of nonlocal contacts which are the same as in the "native" structure, and $N_{\text{total}} = 28$, the number of contacts in any fully compact self-avoiding conformation.

A universal feature of the folding process is that it occurs in two stages (see Fig. 2). There is a rapid compactization ($\sim10^5$ MC steps) in which the number of contacts and energy reach (75–90)% of the values at the global minimum followed by a slow search through the compact conformations for the global minimum. In the rapid compactization process the overlap parameter Q reaches values in the range 0.2–0.4. Then Q fluctuates within this range for a long time (10^6–10^7 MC steps), before it finally reaches the native state ($Q = 1$).

The chain does not encounter significant barriers in the compactization process. To demonstrate this we construct a histogram of the probability of the number of

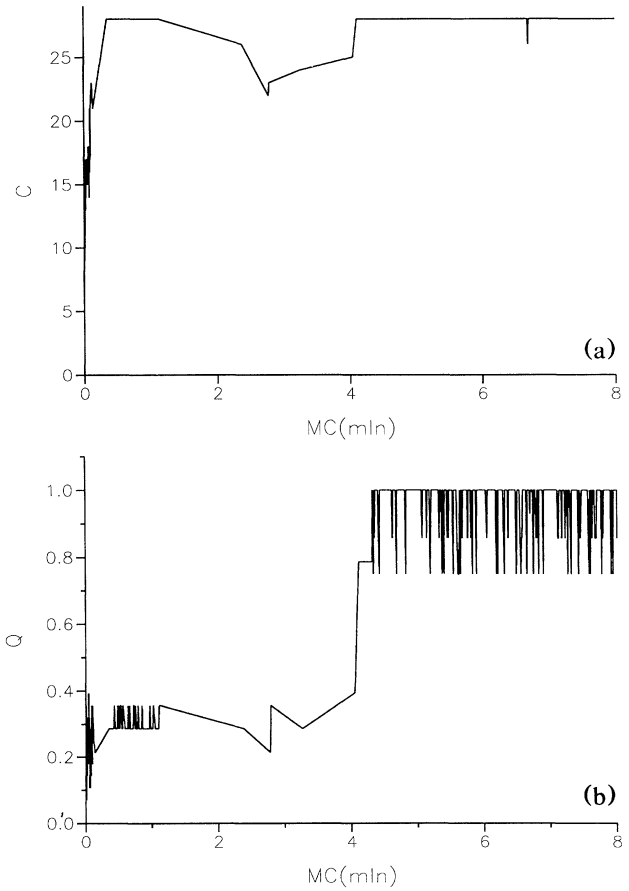

FIG. 2. The MC time evolution of (a) number of nonlocal contacts in the chain and (b) overlap with the ground state for a typical run. The number of nonlocal contacts in CSA is 28.

contacts observed in the course of simulations. The histogram constructed at the temperature which is the midpoint of the transition ($T=3$, Fig. 3) is shown. The most important feature of this histogram is that it is not bimodal; i.e., there is no density (number of contacts) range that corresponds to a free-energy maximum. Thus, the compactization transition is not first order, which is in accord with the theoretical prediction that the coil-globule transition in heteropolymers is second order [9,11]. The bimodality at $T=1$ corresponds to the second transition connected with the search for the ground state (see below).

The search for the ground state is the second stage of folding. It occurs within compact or nearly compact states. In the plot of Q versus the number of MC steps illustrated for a typical run in Fig. 2(b), the chain remains in the space of compact "misfolded" states for a relatively long time ($\sim 10^6$ MC steps) until it suddenly jumps ($\leq 10^4$ MC steps) to a more nearly native configuration ($Q\sim 0.8$), before it finally reaches the ground state. In 36 out of the 50 trials, an intermediate which is a CSA conformation with $Q=0.753$ and $E=-81.02$ was encountered before the native state. In the remaining runs, the second stage ($\leq 10^4$ MC steps) involved a direct transition from $Q\sim 0.2$–0.4 to $Q=1$. In any case there was no unique pathway in a mechanistic sense as a sequence of atomic-scale events. In contrast to the compactization transition, there are significant free-energy barriers separating the various species, e.g., the regions $Q\sim 0.5$–0.7 and $Q\sim 0.8$–0.9 have low probability. Such a first-order transition for formation of a well-defined structure was predicted in analytical investigation of the behavior of random heteropolymers [9].

That the formation of the ground state at equilibrium involves a first-order transition has kinetic implications; i.e., since there are two thermodynamic states separated by a free-energy barrier, there must be a transition region. Nuclei with $Q\cong 0.6$ form the transition state; they have somewhat different structures in the various runs [14]. Sometimes the nucleus dissolves and it takes more time to form a new nucleus from which the ground state is attained.

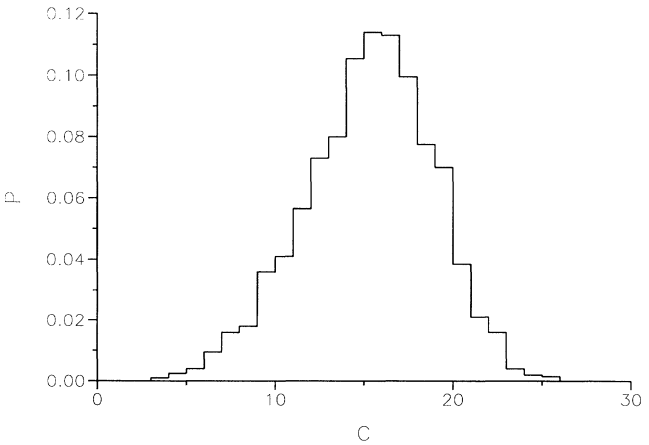

FIG. 3. Histogram for the number of contacts observed in simulations at temperature $T=3.0$.

The most important result is that of thirty sequences examined, only one was able to reach the global minimum in a finite time. Further, this folding sequence did so consistently, independent of the initial conditions. Thus, although the model has a conformation space that exhibits the Levinthal paradox (the total number of conformations of a chain is $6^{25}\approx 10^{20}$), certain sequences can find the global minimum in 10^6–10^7 MC steps during which the chain explores only a negligible fraction ($\approx 10^{-13}$) of the total conformation space. The system also searched only a negligibly small fraction of the compact states. This suggests that, even if the native conformation is a free-energy minimum, there may be a requirement that the conformation be accessible kinetically in a reasonable time. Thus, in addition to selecting sequences in terms of the thermodynamic requirements for a unique global minimum [15], evolution may also have to select sequences in terms of the kinetic requirements for folding. One possibility of such a selection was suggested in [16] where an intrinsic propensity for each monomer to be in native conformation was assumed. It was shown in [17] that under these assumptions the kinetics of folding to the special free-energy minimum may be fast.

To examine the difference between this sequence and other sequences which did fold in the course of simulation, we examine the free-energy landscape. It has been shown that a typical free-energy surface of a random heteropolymer corresponds to a "rugged" landscape; i.e., there are many deep minima with structures significantly different from the native state separated by high barriers [9]. The multiple-minima character of the conformational space of a protein was observed experimentally [18]. A useful measure of the ruggedness of the configuration space of a protein is provided by the function

$$P(Q)=\sum_{kk'}\delta(Q_{kk'}-Q)p_k p_{k'}, \tag{4}$$

where $Q_{kk'}$ is the overlap [Eq. (3)] between conformations k and k'. This function characterizes how low-energy conformations differ structurally; it was calculated analytically for a random heteropolymer in [9]. The summation in Eq. (4) is taken over all conformations, and $p_k=\exp(-E_k/k_BT)/Z$ is the thermal probability for the chain to be in conformation k at a temperature T, E_k is the energy of a chain in conformation k, and Z is the partition function of the chain. The "typical" $P(Q)$ for a random heteropolymer is bimodal with peaks at $Q=1$ and $Q\ll 1$ [9]. Since the search for the native conformation involves the compact configurations (see above), we evaluated $P(Q)$ for the CSA conformations using the method developed in Ref. [6]. Figures 4(a) and 4(b), re-

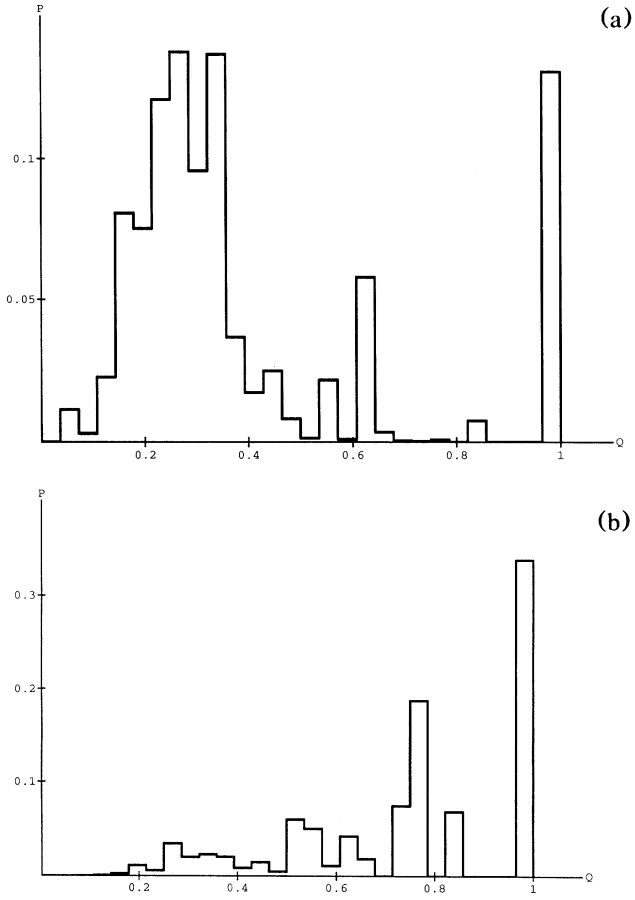

FIG. 4. Plots of function $P(Q)$ [Eq. (4)] at $T=1$. (a) Typical "not-folding sequence" and (b) "folding sequence."

spectively, show $P(Q)$ for a particular sequence that does not fold and for the folding sequence. The configuration space of the "nonfolding" sequence is typical of that for a random heteropolymer. It is rugged with conformations close to the ground state in energy (i.e., highly probable conformations) very different from it structurally. In the process of folding these incorrect conformations serve as "traps" for the chain; i.e., they correspond to deep local minima which in the conformation space are far from the ground state and separated from it by high barriers. By contrast, the configuration space of the folding sequence [Fig. 4(b)] is much smoother. The peak at $0.7 \leq Q \leq 0.8$ corresponds to the CSA conformation which is structurally close to the ground state (it has 22 contacts in common with the ground state out of a total of 28) and energetically close (it is only $1kT$ higher in energy than the ground state). This conformation was often observed as a kinetic intermediate in the second stage of folding before the polymer chain reached the ground state.

It remains to be answered how special the "kinetically folding" sequences are and what fraction of random sequences [15] would meet both the thermodynamic and kinetic requirements of folding. Structural as well as energetic analyses of the difference between folding and nonfolding sequences should help to elucidate this problem.

We would like to thank O. B. Ptitsyn and P. G. Wolynes for fruitful discussions. This work was supported in part by grants from the National Institutes of Health and the National Science Foundation.

[a]Permanent address: Institute of Protein Research, Academy of Sciences of the U.S.S.R., 142292 Puschino Moscow Region, U.S.S.R.

[1] C. Levinthal, J. Chim. Phys. **65**, 44 (1968).
[2] M. Karplus and D. Weaver, Nature (London) **260**, 404 (1976).
[3] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[4] M. Levitt and A. Warshel, Nature (London) **253**, 694 (1975).
[5] C. Wilson and S. Doniach, Proteins: Struct. Funct. Genet. **6**, 193 (1989).
[6] G. Iori, E. Marinari, and G. Parisi, report, 1991 (to be published).
[7] J. D. Honeycutt and D. Thirumalai, report, 1991 (to be published).
[8] E. I. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[9] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[10] A. Yu. Grosberg, S. K. Nechaev, and E. I. Shakhnovich, J. Phys. (Paris) **49**, 2095 (1988).
[11] E. I. Shakhnovich and A. Yu. Grosberg, Zh. Eksp. Teor. Fiz. **91**, 837 (1986) [Sov. Phys. JETP **64**, 493 (1986)].
[12] N. Metropolis *et al.*, J. Chem. Phys. **21**, 1087 (1953).
[13] P. H. Verdier, J. Chem. Phys. **59**, 611 (1973).
[14] T. E. Creighton, Proc. Natl. Acad. Sci. U.S.A. **85**, 5082 (1988).
[15] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).
[16] J. Bryngelson and P. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[17] J. Bryngelson and P. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[18] H. Frauenfelder, F. Parak, and R. Young, Annu. Rev. Biophys. Biophys. Chem. **17**, 451 (1988).