Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 1282–1286, February 1995
Biophysics

Evolution-like selection of fast-folding model proteins

A. M. GUTIN, V. I. ABKEVICH, AND E. I. SHAKHNOVICH

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

Communicated by William Klemperer, Harvard University, Cambridge, MA, November 21, 1994

ABSTRACT **We propose an algorithm providing sequences of model proteins with rapid folding into a given target (native) conformation. This algorithm is applied to a chain of 27 residues on a cubic lattice. It generates sequences with folding 2 orders of magnitude faster than that of the practically random starting sequence. Thermodynamic analysis shows that the increase in speed is matched by an increase in stability: the evolved sequences are much more stable in their native conformation than the initial random sequence. The unfolding temperature for evolved sequences is slightly higher than the simulation temperature, bearing direct correspondence to the relatively low stability of real proteins.**

Biological activity of globular proteins is closely related to the existence of their unique three-dimensional structures. Proteins having unique structure must satisfy two conditions: (*i*) the structure must be thermodynamically stable and (*ii*) it must be kinetically reachable in a biologically reasonable time. It was conjectured by Levinthal (1) that only evolutionarily selected sequences are able to fulfill both of these requirements.

The subsequent development of protein folding theory provided a solid support to this hypothesis. It was shown analytically (2, 3) and numerically (3–6) for several models of proteins that condition *i* of thermodynamic stability is not too restrictive. It was conjectured (2, 3) that there exists a critical temperature, T_c, such that a significant fraction of random sequences have a stable unique structure at $T < T_c$. It was also pointed out (2) that the probability for a random sequence to have a stable unique structure at $T < T_c$ does not depend on its length N. This result was confirmed in a more recent numerical study (6). However, most importantly, the native conformation becomes kinetically inaccessible at $T < T_c$. This was shown first by Bryngelson and Wolynes (7), who found that T_c (T_g in their notation) is a glass transition temperature below which kinetics slows dramatically, so that it takes exponentially long in chain-length N time ("Levinthal" time) to reach the ground state at $T < T_c$. This introduces the new paradox: for random sequences, $T < T_c$ for stability and $T > T_c$ for kinetic accessibility of the native state. The implication is that random sequences are not able to fold into a unique structure.

The principal way out is to find special sequences that have their native structures stable at $T > T_c$ (8), resolving the contradiction between the thermodynamic and kinetic requirements characteristic for random sequences. The requirement that a sequence has a native structure stable at $T > T_c$ imposes the necessary condition that it is a pronounced energy minimum separated by a large energy gap from the set of nonnative conformations (8–10). To meet the requirement of chain stability at $T > T_c$, independent of chain length, the size of the energy gap Δ must scale with the chain length N as $\Delta \sim Nk_BT_c$ (8, 9, 11). The probability P_Δ of finding a random sequence with energy gap Δ was estimated (3, 10) to be $\sim\exp(-\Delta/k_BT_c)$. This implies that only an exponentially vanishing fraction of random sequences [$\sim\exp(-\alpha N)$] are able to fold. In contrast, the energy gap for a typical random sequence is of the order of a few k_BT_c for any chain length; that is why a sufficiently low temperature is needed to make their native state stable (2).

The probability of finding a folding sequence by simply pulling it out at random from the "soup" of all possible sequences is very small for chains of realistic length. This seems to introduce a "Levinthal paradox" in sequence space. However, this difficulty may be overcome by a simple sequence design algorithm suggested in refs. 9 and 12. The algorithm generates sequences that have sufficiently low energy in a chosen target conformation. The energy gap in designed sequences was indeed large enough to enable them to fold fast to the stable native conformation (13–15), which coincided with the target conformation used at the design stage. The design algorithm is based on the idea of Monte Carlo search in sequence space. It proceeds by making mutations in sequences, biasing them by Metropolis criterion (16) with "selective temperature" T_{sel} to sequences with low energy of the native state. The Metropolis bias in the algorithm made it possible to find sequences having energies sufficiently low for folding in spite of the fact that the fraction of such sequences among all possible ones is exponentially small, although their number is still exponentially large.

The results of refs. 9, 14, and 15 demonstrate that sequence design aimed at generating thermodynamically stable sequences for a specific conformation makes this conformation accessible—i.e., that thermodynamic stability is a sufficient condition for fast folding. We may ask now, Is it also a necessary condition for fast folding?

To address this question, we invert the approach taken in refs. 9 and 14 and develop a design algorithm aimed at direct optimization of the folding rate to provide fast-folding sequences. The idea of the algorithm is simple. An attempt at a random point mutation is made. If the new sequence folds slower than the current sequence, then the mutation is rejected with large probability. If the mutation results in faster folding, then with high probability the mutation is accepted. Therefore, after a large enough number of mutations we expect to generate fast-folding sequences.

The idea of our selection algorithm is similar in spirit to the idea of simulated annealing. However, there is an important difference. Usually, in simulated annealing the optimized quantity (for example, the energy of a system) can be calculated exactly. This is not the case for the selection algorithm. In this case, the optimized quantity is the folding time or, more precisely, the mean first passage time (MFPT) to the native conformation. An exact determination of this quantity would require an infinite number of folding simulations. In reality, we can have only an estimate of the MFPT. Moreover, each folding takes a long time; therefore, we cannot afford to get very good statistics.

Taking this into account, we developed the following three-step algorithm for evaluation of the kinetic consequence of a point mutation. In step 1, we perform two folding runs and estimate very roughly the new MFPT. If it is longer than the original MFPT, then the mutation is rejected; if the new MFPT is shorter, then we estimate the new MFPT more precisely in step

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

Abbreviation: MFPT, mean first passage time.

2. The purpose of step 1 is to reject outright obviously poor mutations, which constitute the majority of all mutations. In step 2, we perform 10 folding simulations and therefore get a much more precise estimate for the MFPT. If it is less than the original MFPT by 20%, then the mutation is accepted; otherwise, it is rejected. If the mutation is accepted, then an additional 100 folding runs are performed (step 3) to get a reasonably good estimate for the new MFPT. This estimate is used as a current MFPT for comparisons with the MFPT of the next mutations.

The consistent implementation of the idea of searching for fast-folding sequences requires that the algorithm starts from a random sequence and then scans sequence space for fast-folding sequences as described above. The important requirement here is that the algorithm should be seeded with an initial random sequence, which folds in a reasonable amount of computer time. This requirement prohibits the use of long sequences for this study. Although we can fold long designed sequences, up to 100 residues (13, 14), folding of long random sequences requires Levinthal time (13, 14). Therefore, with long chains we cannot provide a starting sequence for the selection algorithm. A significant fraction of short random chains (27 residues) can fold (10). This restricts our study to chains of 27 monomers on a cubic lattice (2, 4, 9, 10, 17, 18). The energy of a conformation of the chain is the sum of energies of pairwise contacts:

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

where $\Delta_{ij} = 1$ if monomers i and j are lattice neighbors and $\Delta_{ij} = 0$ otherwise. ξ_i defines the type of amino acid residue in position i. $B(\xi, \eta)$ is a magnitude of contact interaction between amino acids of types ξ and η. We expect that the choice of the specific set of parameters for our study is not very essential (for more detailed discussion of this problem, see ref. 15). We used the parameter set published (19). This set of parameters was derived from statistics of contacts in proteins in quasi-chemical approximation; it serves our purposes as a set of uniformly distributed numbers. B_0 is an energy parameter having the meaning of overall attraction; as in previous works (10, 17), it was introduced to bias conformations toward more compact ones, forcing the native state to belong to the set of maximally compact conformations. The motion of the chain is modeled by the standard cubic lattice Monte Carlo algorithm (20, 21). Simulations were performed at temperature $T = 0.32$ and with $B_0 = -T = -0.32$.*

First, we generated 10 random sequences. Then, by exhaustive enumeration (4), we found the maximally compact conformation with the lowest energy for each of these sequences. Then, we picked the sequence (Fig. 1*A*) that folded to its native conformation (Fig. 1*B*). The MFPT for the sequence shown in Fig. 1*A* at temperature $T = 0.32$ is close to 5×10^6 Monte Carlo steps. This sequence was used to start the selection.

The result of the selection is presented in Fig. 2, which shows the MFPT as a function of the number of accepted mutations. It can be seen that for ≈100 accepted mutations the MFPT decreased almost 2 orders of magnitude. To give an idea of how the selection algorithm works, we note that 300 accepted mutations presented in Fig. 2 correspond to ≈10,000 attempted mutations, 3000 of which were accepted at step 1 and passed to step 2 of the algorithm. The simulation required ≈100 hr of CPU time on the IBM RS6000 computer.

*The reader may notice the difference in temperature scales between the present work and previous lattice model simulations from our group (10, 14, 15, 18, 22). This is due to the fact that, in previous studies, we scaled the parameter set to have $\langle B^2 \rangle = 1$, while in this study we use parameters directly as published (19). The difference is only a constant factor in temperature scales.

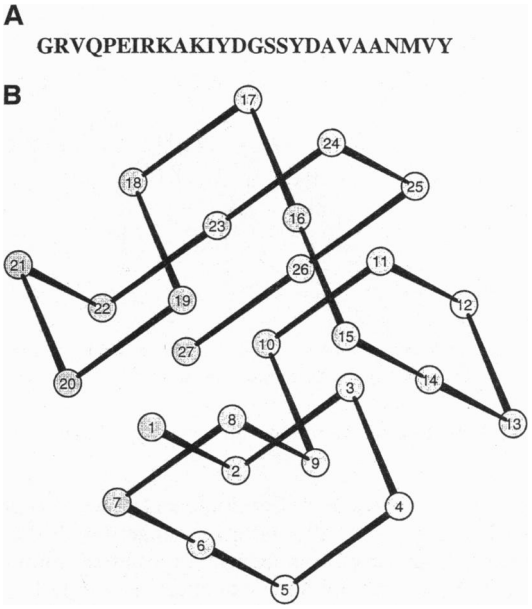

FIG. 1. A starting quasi-random sequence (*A*) and its maximally compact conformation with lowest energy (*B*).

One way to make a chain fold rapidly is to make the energy E_{nat} of its native conformation as low as possible. The absolute value of the energy itself is not directly related to stability; what is important is the relative value of energy, or Z score, introduced by Eisenberg and coauthors (23):

$$Z = \frac{E_{nat} - E_{av}}{\sigma}, \qquad [2]$$

where E_{av} is the average energy of compact nonnative conformations. To estimate this value, we determined all topologically possible contacts between all monomers. Then, we calculated the average energy e_{av} of a contact and the corresponding dispersion σ. The average energy of nonnative conformation can be estimated as $E_{av} = Ke_{av}$, where $K = 28$ is the number of contacts in a maximally compact conformation.

The evolution of the relative energy of the native conformation Z is shown in Fig. 3. It is seen that for the first 50 accepted mutations, when the MFPT decreases, Z decreases noticeably too, although because of strong fluctuations of Z the effect is not very pronounced. To get a clear impression of the properties of the ensemble of selected fast-folding sequences, we calculated the distribution of Z over 250 sequences starting from the 50th accepted mutation. The corresponding

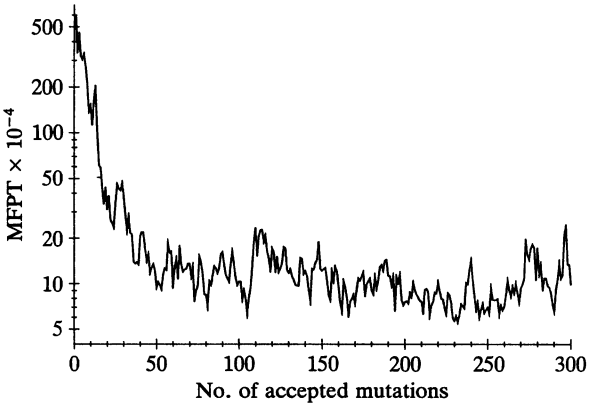

FIG. 2. Evolution of the MFPT (in Monte Carlo steps).

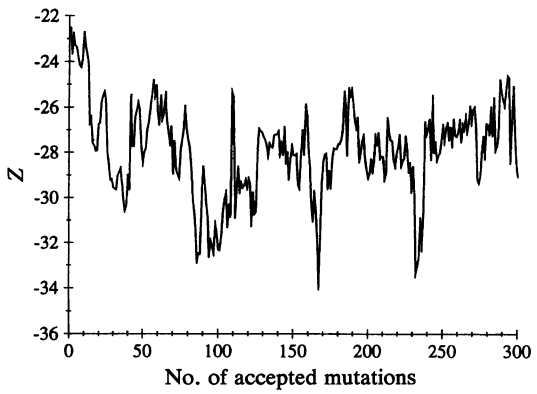

FIG. 3. Evolution of the relative energy of the native conformation Z.

histogram is shown in Fig. 4. For comparison, we also plotted a similar histogram for 250 random sequences with the same amino acid composition. It is seen that the distribution of the parameter Z for fast-folding sequences generated by the selection algorithm is shifted to more negative values compared to random sequences; the most probable value of Z for fast-folding sequences is about -27.5 and for random sequences it is about -20. The distribution of Z for random sequences can be fit very well by a Gaussian distribution. It is possible to determine from such a fit the probability to have a random sequence with $Z = -27.5$ corresponding to the median of the Z distribution for selected sequences. The elementary estimate gives $P(Z = -27.5) \approx 10^{-6}$. Therefore, 1 of 10^6 random sequences for the 27-mer will behave like an average fast-folding sequence from the pool of "evolutionarily selected" ones. We also emphasize that this estimate strongly depends on chain length, and for longer sequences this probability will be much lower.

The parameter Z estimates the energy of the native conformation relative to other conformations. There is a linear dependence between Z and the energy gap discussed in previous publications from several groups (2, 8, 9, 10, 14, 24). A more direct way to analyze the thermodynamic stability of the native state is to estimate its unfolding temperature T_f (8, 14, 25). T_f is defined as temperature of midtransition between the native and denatured state. To determine T_f, it is convenient to use the parameter Q, which is the average at a given temperature of the number of native contacts (14). The native conformation has $Q = 28$, while a typical nonnative conformation has only a few native contacts. The temperature dependence of Q exhibits a sigmoidal shape (see, e.g., figure 4 of ref. 22). We define transition temperature T_f as a temperature at which $Q = 14$.

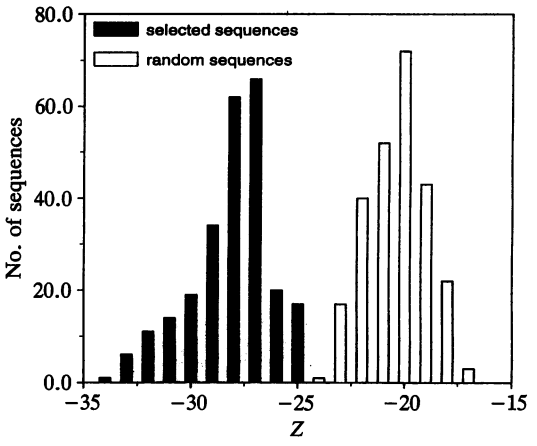

FIG. 4. Distribution of the relative energy of the native conformation Z for random sequences (open bars) and for fast-folding sequences generated by the selection algorithm (solid bars).

Similar plots were obtained for some sequences generated by the selection algorithm. The corresponding unfolding temperatures are presented in Fig. 5. For the first 50 mutations, when the MFPT decreases, the unfolding temperature increases, which is equivalent to an increase of stability of the native structure. For fast-folding sequences (after 50 accepted mutations, when the curves in Figs. 2 and 3 saturate) the unfolding temperature T_f appears to be only slightly higher than the temperature $T = 0.32$ at which the simulations were done.

It was shown in a number of previous calculations that thermodynamic stabilization results in rapid folding. In this paper, we have shown that the opposite is true as well: optimization of folding rate results in a pronounced thermodynamic stabilization of the native state.

The physical explanation of correlation between speed and stability is made clear if we compare the density of states of random and selected, fast-folding sequences, which have a large Z score. The two important parameters must be taken into account in the description of configurational space of a chain. These are the energy of the chain E and the degree of folding Q, defined as the number of contacts that are the same as in the native conformation (4, 14, 18). The quantity of interest here is the logarithm of the density of states $\nu(E, Q)$ (18). The evaluation of this quantity is based on the "histogram method," which assumes that a Boltzmann distribution is achieved in the process of Monte Carlo simulation. The key feature that allows us to determine $\nu(E, Q)$ unambiguously is the fact that the ground (native) conformation is nondegenerate and its energy is known. It gives the base point to evaluate densities of states for any energy and degree of folding observed in simulations. We plot in Fig. 6 these quantities for three cases: for a random sequence, for a "typical" evolved sequence, 300, having $Z \approx -29$, and for one of the most stable and fast-folding evolved sequences, 167, with $Z \approx -34$. One can see clearly why optimization of the energy of the native state is required for fast folding and how it eliminates the multiple-minima problem. Fig. 6 also clarifies the concept of the energy gap, which caused some controversy. The gap, which is relevant for folding, is the energy difference between the native state and lowest energy misfolded conformations. One can clearly see the pronounced difference in this parameter for the random sequence (Fig. 6*A*) and evolved fast-folding sequences (Fig. 6 *B* and *C*). We define as misfolded the lowest energy conformation having only five native contacts (the degree of similarity expected for two randomly superimposed compact conformations).

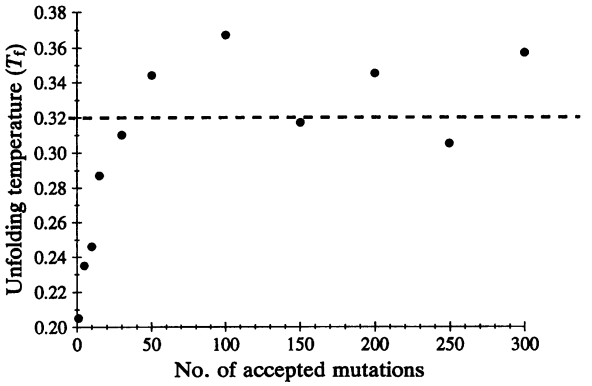

FIG. 5. Evolution of the unfolding temperature T_f. Dashed line denotes simulation temperature $T = 0.32$.

For random sequence, the relative value of the gap $\Delta' = (E_{native} - E_{misfolded})/E_{native} \approx 0.1$, while for sequence 300 (Fig. 6*B*) $\Delta' \approx 0.23$ and for sequence 167 (Fig. 6*C*) $\Delta' \approx 0.31$. Fig. 6 shows how optimization creates a driving force to the native state; decrease of energy for evolved sequences is contingent on coming close to the native state, while for random sequences a multitude of misfolded conformations have almost the same energy as the native state.

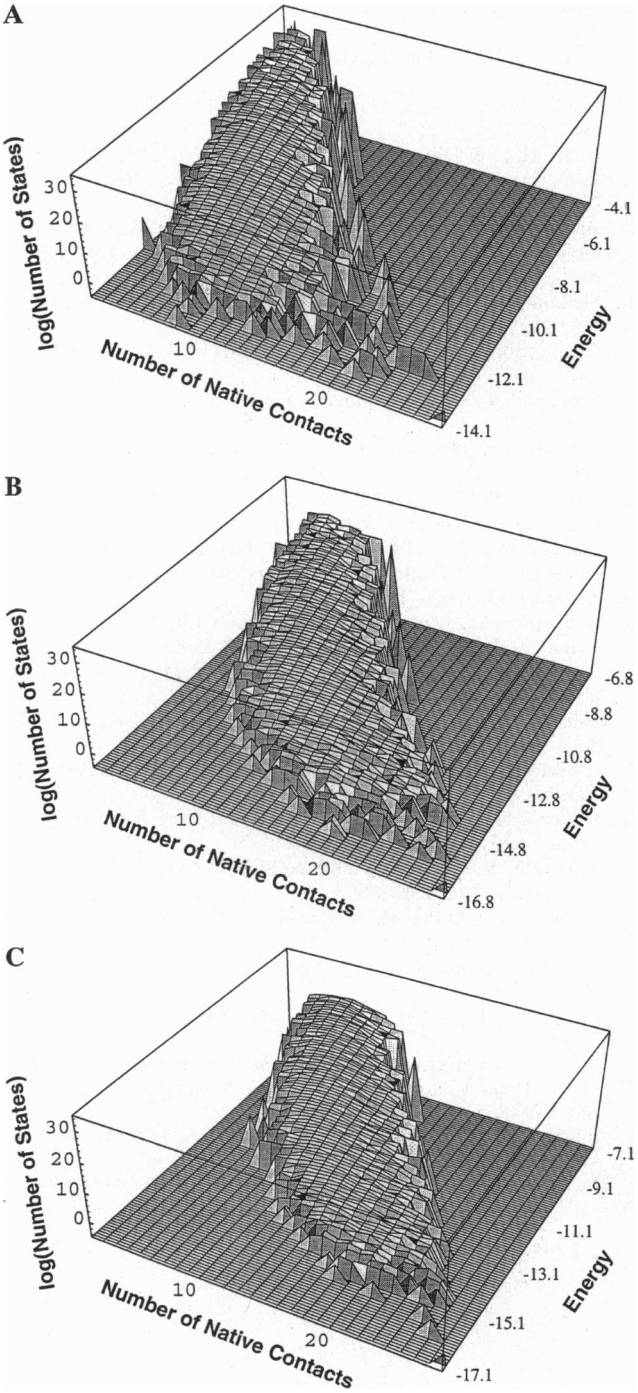

FIG. 6. Density of states (logarithmic scale) for a random sequence (*A*), the last evolved, 300 (*B*), and the best evolved, 167 (*C*), plotted as a function of energy and degree of folding, which is the number of native contacts. Random sequence folds in $\approx 5 \times 10^6$ Monte Carlo steps, sequence 300 folds in ≈140,000 Monte Carlo steps, and sequence 167 folds in ≈70,000 Monte Carlo steps. Their Z scores are approximately −20, −29, and −34, respectively. For each sequence, a long equilibrium Monte Carlo simulation was run, and the statistics of occurrence of different states were collected. The procedure of how to obtain the density of states from such statistics is described (18).

This is in accord with earlier findings that sequences having lower energy in the native conformation generally fold faster (9, 10, 22, 26). At the same time, the correlation between folding rate and stability is pronounced only at relatively high temperature (22). As temperature decreases, approaching T_c, the MFPT becomes independent of the energy of the native conformation. This is explained by the fact that at low temperature a chain often first is trapped in a low energy misfolded conformation, and the MFPT is determined mainly by the time necessary to escape from the trap, which should not depend on the energy of the native conformation (7). To understand what happens to the selection algorithm at low temperature, we ran the algorithm at temperature $T = 0.2$. It appeared that, in this case, the algorithm also was able to make folding many times faster, although the selected sequences had the MFPT larger than those generated at $T = 0.32$. How did the algorithm manage to make folding faster?

Since at low temperatures the MFPT is determined by the time necessary to escape from low energy traps, the simplest imaginable way to decrease the MFPT would be to destabilize the traps. This is done by increasing the average energy of a contact e_{av} or by decreasing the corresponding dispersion σ (Eq. **2**). This is what we observed at the low temperature run of the selection algorithm.

We already mentioned that the selection algorithm is similar to standard simulated annealing. There is also a loose similarity with biological evolution: random mutations and selection pressure. In this sense, the finding that selected sequences have their middenaturation transition temperature T_f close to the temperature of folding simulations may be relevant. This is in accord with the fact that most of the globular proteins are not very stable; the free energy difference between the native and the denatured states is relatively small (25). The set of sequences generated by the selection algorithm also contains a significant proportion of more stable ones with $Z < -30$. If additional selective pressure toward greater stability is applied, then more stable sequences from the generated pool can be picked up easily since their proportion is significant (>20%; see Fig. 4).

In this work, we emphasize that there exists a correlation between speed and stability. However, our results also indicate that this correlation may be limited as well, and there are factors other than overall stability to be taken into account in analyzing folding rates. Indeed, we can compare the plots in Fig. 3 and Fig. 2 to see that there are pronounced fluctuations in the parameter Z in the steady-state part of the plot in Fig. 3 that are not matched by equally pronounced fluctuations in folding rate in Fig. 2. This suggests that at any Z there is a distribution of sequences having different folding rates. This fact was reported in an earlier publication (22). The observed correlation between Z and the folding rate implies that the distribution of folding rates for sequences having smaller Z is shifted toward faster folding compared to that for sequences having higher Z. Our algorithm, optimizing for higher folding rates, selected sequences that at any given Z belong to the fast-folding tail of this distribution. It is interesting in this regard to note asymmetry of the Z histogram for the selected sequences presented in Fig. 4. Selection pressure from the algorithm results in a fast decay of the distribution at large Z; the algorithm does not find fast-folding sequences with $Z > -26$. The slow decay of the distribution at low Z can be fit by an exponential. This is likely due to the interplay of two factors (as in any optimization algorithm): bias toward further optimization and the opposing entropic factor. In our case, this means that although it is more likely to find a fast-folding sequence among low Z sequences, there are too few of them, so that it becomes unlikely that the algorithm finds such sequences. The steady state seen in Figs. 2 and 3 is reached due

to the balance between the tendency to optimize the folding rate and entropy in sequence space. This stabilizes the histogram in Fig. 4, making it peak around $Z \approx -28$ with statistical fluctuations toward lower Z values. The parameter that governs the balance between the optimized parameter and entropy in simulated annealing is temperature. In our case, the effective temperature plays the role of the criterion of acceptance of a mutation at step 2 (20% increase in rate). Decrease of temperature is equivalent to a more demanding requirement that the rate increases upon accepted mutation, but like any optimization algorithm this one will freeze at lower temperature because of insurmountable barriers in sequence space.

It was found experimentally (see table I of ref. 27 and table II of ref. 28) and in lattice model simulations (15) that stability is not a single factor determining the folding rate: various mutations may have a comparable impact on stability while having a differing impact on folding rate. To the best of our knowledge, the only folding mechanism consistent with these experimental findings, where there is a clear distinction between "kinetically important" residues and other residues, is the nucleation growth mechanism via a specific nucleus (15). It was shown (15) that sequences have different folding rates while having the same energy in the native state. The important factor is how this stabilization energy is distributed between nucleus contacts and the remaining ones. We cannot exclude the possibility that there may be other factors determining folding rates. Our approach provides a unique opportunity to study these factors, providing numerous sequences, which were selected as fast-folding ones, and comparing them with a number of other sequences selected by thermodynamic design (9, 12) to have the same Z score. The statistically meaningful differences in their folding rate point to features, in addition to thermodynamic stabilization, that lead to faster folding.

In this paper, we have applied the selection algorithm to the simple lattice model of proteins analogous to ones used in previous studies (4, 5, 10, 17). Lattice models of proteins proved useful for investigation of principal aspects of protein folding, especially its faster stages when backbone topology becomes established. It is almost certain that more detailed models are required to study the final stages of folding where the native structure forms and side-chain packing becomes important. There are cases when fast, two-state, kinetics describes the whole folding process, up to formation of the native protein (29). It is conceivable that tight packing of side chains is less important for kinetics here and therefore lattice model simulations may be more applicable to such cases.

In conclusion, a simple evolution-like algorithm proposed in this paper generated fast-folding sequences. All of them have low relative energy of the native conformation compared to random sequences, and all are considerably more stable in the native conformation at the simulation temperature. This implies that lowering the native state energy is a necessary condition for fast folding in the studied model of proteins.

This work was supported by a Packard Fellowship.

1. Levinthal, C. (1969) *Mossbauer Spectroscopy of Biological System* (Univ. of Illinois Press, Urbana).
2. Shakhnovich, E. I. & Gutin, A. M. (1990) *Nature (London)* **346,** 773–775.
3. Gutin, A. M. & Shakhnovich, E. I. (1993) *J. Chem. Phys.* **98,** 8174–8177.
4. Shakhnovich, E. I. & Gutin, A. M. (1990) *J. Chem. Phys.* **93,** 5967–5971.
5. Chan, H. S. & Dill, K. A. (1991) *J. Chem. Phys.* **95,** 3775–3787.
6. Camacho, C. & Thirumalai, D. (1993) *Phys. Rev. Lett.* **71,** 2505–2508.
7. Bryngelson, J. D. & Wolynes, P. (1989) *J. Phys. Chem.* **93,** 6902.
8. Goldstein, R., Luthey-Schulten, Z. A. & Wolynes, P. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9029–9033.
9. Shakhnovich, E. I. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7195–7199.
10. Sali, A., Shakhnovich, E. I. & Karplus, M. (1994) *J. Mol. Biol.* **235,** 1614–1636.
11. Ramanathan, S. & Shakhnovich, E. (1994) *Phys. Rev. Sect. E.* **50,** 1303–1312.
12. Shakhnovich, E. I. & Gutin, A. M. (1993) *Protein Eng.* **6,** 793–800.
13. Shakhnovich, E. (1994) *Protein Structure by Distance Analysis* (IOS, Amsterdam), pp. 201–212.
14. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72,** 3907–3909.
15. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *Biochemistry* **33,** 10026–10037.
16. Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, E. & Teller, J. (1953) *J. Chem. Phys.* **21,** 1087–1099.
17. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991) *Phys. Rev. Lett.* **67,** 1665–1667.
18. Sali, A., Shakhnovich, E. I. & Karplus, M. (1994) *Nature (London)* **369,** 248–251.
19. Myazawa, S. & Jernigan, R. (1985) *Macromolecules* **18,** 534–552.
20. Verdier, P. H. (1973) *J. Chem. Phys.* **59,** 6119.
21. Hilhorst, H. J. & Deutch, J. M. (1975) *J. Chem. Phys.* **63,** 5153–5161.
22. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *J. Chem. Phys.* **101,** 5062–6062.
23. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991) *Science* **253,** 164–169.
24. Guo, Z., Thirumalai, D. & Honeycutt, J. D. (1992) *J. Chem. Phys.* **97,** 525–535.
25. Privalov, P. L. (1989) *Annu. Rev. Biophys. Biophys. Chem.* **18,** 47–69.
26. Socci, N. & Onuchic, J. (1994) *J. Chem. Phys.* **101,** 1519–1528.
27. Matouschek, A., Serrano, L. & Fersht, A. R. (1992) *J. Mol. Biol.* **224,** 819–835.
28. Jackson, S., elMasry, N. & Fersht, A. R. (1993) *Biochemistry* **32,** 11270–11278.
29. Jackson, E. & Fersht, A. (1991) *Biochemistry* **30,** 10428–10435.