

Conformation, energy, and folding ability of selected amino acid sequences

(protein folding/inverse folding/sequence design)

M. SASAI

Graduate School of Human Informatics, Nagoya University, Nagoya 464-01, Japan

Communicated by Peter G. Wolynes, University of Illinois, Urbana, IL, April 21, 1995

ABSTRACT Evolutionary selection of sequences is studied with a knowledge-based Hamiltonian to find the design principle for folding to a model protein structure. With sequences selected by naive energy minimization, the model structure tends to be unstable and the folding ability is low. Sequences with high folding ability have not only the low-lying energy minimum but also an energy landscape which is similar to that found for the native sequence over a wide region of the conformation space. Though there is a large fluctuation in foldable sequences, the hydrophobicity pattern and the glycine locations are preserved among them. Implications of the design principle for the molecular mechanism of folding are discussed.

A protein chain can take on an enormous number of different conformations, and an astronomically long time is required for an exhaustive survey of those conformations. How does a protein accomplish the fast structural search to find its unique native structure (1)? One possible explanation is that the native structure corresponds to a pronounced minimum of free energy, so that the thermodynamic predominance should assure its kinetic preference at the same time. This mechanism was discussed by Bryngelson and Wolynes (2, 3) and was quantitatively analyzed with replica methods (4–7). More recently, thorough investigations have been done with lattice models (8–14): Leopold *et al.* (8) showed that convergent pathways, or “funnels,” guided folding to the native structure; Dill and colleagues (11, 12) suggested that “hydrophobic zipper” processes were responsible; and Sali *et al.* (13) showed that the chain should fold efficiently when the energy of the correct structure was low enough. All these theoretical results support the picture that the energy landscape of the foldable chain must have some coherence which leads to the convergent pathways to the native structure. From the spin-glass theoretical point of view, the physical basis for this coherence should be the minimal frustration among interacting residues (2, 3, 15).

One way to examine this “minimal frustration” principle and to find the further molecular mechanism of folding is to look by computer for sequences that are compatible with a given three-dimensional structure. This problem is termed the inverse folding problem (16–28). By finding sequences which can fold to the given structure, we could understand the design principle for sequences to have efficient folding abilities. Dill and coworkers (17–19) examined sequences in the lattice model and showed that hydrophobic interactions played the key role. Shakhnovich and Gutin (20, 21) performed artificial evolution experiments by selecting randomly generated sequences. These studies (17–21), however, were based on simplified lattice models, and many important features of real proteins were lacking. Thus, it is strongly desired to examine

sequence-selection experiments with a more realistic model of proteins.

The aim of this paper is to reveal new design principles through simulated evolution of sequences. The amino acid sequence space is randomly sampled and selected according to several different criteria by using an off-lattice knowledge-based Hamiltonian. First, the Hamiltonian is explained and forward-folding results are discussed. Then the random sequence-space walk is explained and results with different selection criteria are compared.

A Knowledge-Based Hamiltonian

We express the energy of the chain as a sum of pairwise potentials. The potential is constructed from a library of 75 protein structures. These 75 structures were selected as a subgroup of the library used in Table 2 of ref. 24. For short sequence distance, when the residues p and q are found at positions i and $i+k$ in structure μ of the library, a Gaussian function whose center is at $r = r_{i,i+k}^{pq,\mu}$ is summed into the potential V_k^{pq} :

$$V_k^{pq}(r) = \frac{-1}{N_k(2\pi c_k)^{1/2}} \sum_{\mu} \sum_i \exp\left[-\frac{(r - r_{i,i+k}^{pq,\mu})^2}{2c_k}\right],$$

for $k \leq 10$, [1]

where p and q represent 2 of the 20 amino acids, N_k is the number of pairs that appear in the library, c_k is chosen to be $0.5k \text{ \AA}^2$, and $r_{i,i+k}^{pq,\mu}$ is the spatial distance between β -carbon (C^β) atoms. Here, C^β is used instead of the α -carbon C^α because C^β is much more sensitive to the conformation (25). For glycine, C^α is used instead of C^β . For longer sequence distance, contributions from residues at distance j with $m(k) \leq j < m(k+1)$ are summed into the class- k potential:

$$V_k^{pq}(r) = \frac{-1}{N_k(2\pi)^{1/2}} \sum_{m(k) \leq j < m(k+1)} \frac{1}{(c_j)^{1/2}} \sum_{\mu} \sum_i \exp\left[-\frac{(r - r_{i,i+j}^{pq,\mu})^2}{2c_j}\right],$$

for $11 \leq k \leq 15$, [2]

where $m(k)$ is chosen as $m(11) = 11$, $m(12) = 31$, $m(13) = 61$, $m(14) = 101$, $m(15) = 151$, and $m(16) = \infty$. Thus, this expression has a close similarity to that of Sippl and coworkers (23, 24). As in Sippl's potential (24), V_k^{pq} has two sharp minima for many pq combinations. One minimum is at the spatial distance in the α -helix and the other corresponds to the β -sheet. The energy depth of these minima depends on the helix or sheet propensities of p and q .

Additional hydrophobic interactions are introduced with similar Gaussian functions.

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.

Abbreviations: MD, molecular dynamics; HLH, helix-loop-helix.

$$\phi_k^{pq}(r) = \frac{1}{(2\pi c_{sa})^{1/2}} \exp\left(\frac{-r^2}{2c_{sa}}\right) - \frac{1}{(2\pi c_h)^{1/2}} \left[\xi_p \xi_q + (1 - \xi_p \xi_q) \exp\left\{-\frac{(r - r_h)^2}{2c_h}\right\} \right], \quad [3]$$

where $\xi_p = 1$ when p is hydrophilic and $\xi_p = 0$ when hydrophobic. The first term is the self-avoiding repulsion and the second term is the shallow attraction potential. Parameters were chosen to stabilize the native conformation: $c_{sa} = 16 \text{ \AA}^2$, $r_h = 4 \text{ \AA}$, and $c_h = 900 \text{ \AA}^2$. Then the energy of the chain is a sum of these 15 class potentials:

$$E = \sum_{i>j} [a_k V_k^{pq}(r_{ij}) + b_k \phi_k^{pq}(r_{ij})], \quad [4]$$

where r_{ij} is the spatial distance between C^β of the i th and j th residues of the protein under consideration and k is chosen to satisfy $m(k) \leq (i - j) < m(k + 1)$ for $k \geq 11$, and $k = i - j$ for $k \leq 10$. To make the contributions balanced, weight factors are chosen to be $a_k = 1$ and $b_k = 0$ for $k \leq 10$ and $a_k = 0.5$ and $b_k = 0.4$ for $k \geq 11$.

Here, the library data are superposed to obtain the smooth potentials suitable for molecular dynamics (MD) calculation. By similar Gaussian summations, potentials for the neural-network models of folding were constructed (29, 30). Their potentials depend on the absolute position of residues in the sequence and thus are context-dependent. The potentials in Eqs. 1 and 2, on the other hand, depend only on the relative sequence distance and are context-independent. We will restrict our discussion to the context-independent form of potentials because it is more straightforward to use them to test many newly generated artificial sequences.

Forward Folding of a Calcium-Binding Protein

One of the simplest four-helix bundle proteins, a Ca^{2+} -binding protein (Protein Data Bank code 3icb; number of residues, $N = 75$), is used as a target structure for the forward-folding simulation. Neither 3icb nor its homologue is included in the library for potentials. Since only the C^β coordinates are taken into account, we consider the virtual polymer chain connecting C^β atoms. Fig. 1a is the C^β -chain representation of the 3icb x-ray structure. It has two helix-loop-helix (HLH) motifs stacked with canted angles. Ca^{2+} ions are bound at the loop positions. Here we focus only on these topological features, although the neglected structural details would be crucial for Ca^{2+} affinity.

Forces acting upon the virtual polymer chain are derived from Eq. 4. First, the 3icb sequence is threaded on the chain. Starting from the x-ray structure, the structure without Ca^{2+}

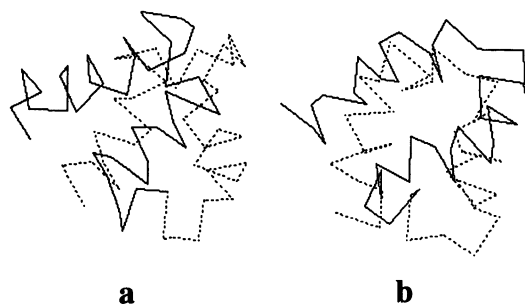


FIG. 1. (a) The x-ray structure of a Ca^{2+} -binding protein (Protein Data Bank code 3icb). C^β atoms are connected by a solid line for residues 1-40 and by a dashed line for residues 41-75. Ca^{2+} ions are bound at the loop positions. (b) The apo structure obtained by the folding simulation has 6.2- \AA rms deviation from the structure in a and 4.8- \AA rms deviation from the reference apo structure.

ions (apo state) is optimized with these forces. This optimized apo structure is used as the reference structure in the following calculations.

Starting from the stretched conformation, the motion of the chain is simulated with a Brownian-motion MD (MD with random forces) calculation. Then, by gradually lowering the noise level with the factor $1/\log(t + t_0)$, the optimal conformation is obtained after 10^5 simulation steps. Here t is the number of steps and t_0 is a constant chosen to be 200. This simulated-annealing calculation was tried several times with different random-number implementation. For 30% of the simulated annealing runs (6 runs out of 20 trial runs), structures with rms difference $< 5 \text{ \AA}$ from the reference structure or from its mirror image were obtained. An example of the structure is shown in Fig. 1b. It has two HLH motifs stacked with the proper orientation. For the other 70% of the runs, two HLH motifs were stacked in the wrong direction and the rms deviation was $> 10 \text{ \AA}$. One reason for this high yield of incorrect conformations would be the lack of dihedral-angle restriction.

The quality of the present potential was examined by threading the 3icb sequence on other structures in the library. Energies were found to spread with width γE around the average value E_{av} . The energy of the reference structure, E_{fold} , is well below E_{av} ; $\Delta E/\gamma E = 4.42$, where $\Delta E = E_{av} - E_{fold}$. The energy of the typical incorrect result, $E_{misfold}$, is well above E_{fold} ; $\eta E/\gamma E \approx 3$, where $\eta E = E_{av} - E_{misfold}$.

Structures that appear along the folding pathway are shown in Fig. 2. The chain folds with three successive steps: 1, formation of helices and primordial HLH motifs; 2, collapse to the globule; and 3, structural search in the globular state. Steps 1 and 2 occur quickly, but a much longer time is needed for step 3. Distinctive minima in the short-range potentials guide the chain to the α -helix formation at an early stage. In step 2, hydrophobic interactions play the decisive role. Since the primordial forms of the second structures have already appeared in step 1, hydrophobic interactions between these structural units in step 2 greatly affect the overall topology. In step 3, structures in various length scales continue to develop, and competitions arise among them. After a long structural search, these structural conflicts are minimized. By changing the random-number implementation, similar pathways are found when the trajectory reaches the correct conformation,

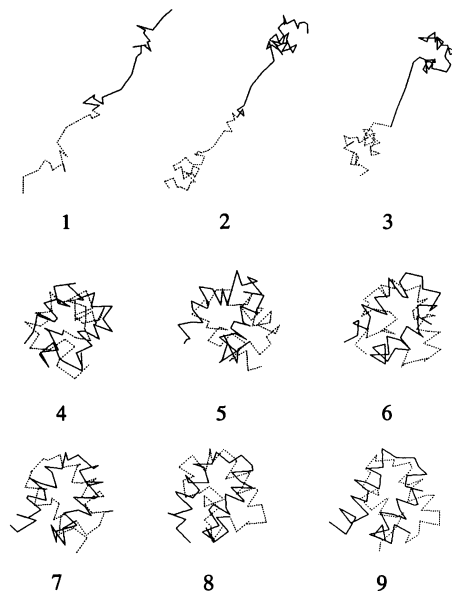


FIG. 2. Structures that appear along the folding pathway are shown at every 2×10^3 steps. Structures 4-9 are drawn with a length scale twice as large as that in structures 1-3.

and the pathways seem to be “funneled” in the sense of Leopold *et al.* (8).

The probability of reaching the correct conformation does not depend much on parameters in Eq. 3 but depends sensitively on a_k and b_k in Eq. 4. When $b_k = 0$ for all k , the probability of finding the correct conformation decreases to a few percent. The probability is maximum when both $\sum_{k \leq 10} a_k V_k^{pq} \approx \sum_{k \geq 11} (a_k V_k^{pq} + b_k \Phi_k^{pq})$ and $\sum_{k \geq 11} a_k V_k^{pq} \approx \sum_{k \geq 11} b_k \Phi_k^{pq}$ are satisfied for the reference structure. When $\sum_{k \leq 10} a_k V_k^{pq}$ is smaller than $\sum_{k \geq 11} (a_k V_k^{pq} + b_k \Phi_k^{pq})$, the chain collapses before the short-range orders develop, and a longer time is needed for the structural search in step 3. When $\sum_{k \leq 10} a_k V_k^{pq}$ is too large, on the other hand, the flexibility of the chain is lost and the probability of folding is low.

When the rms deviation is $< 5 \text{ \AA}$, the simulated structure is regarded to be topologically the same as the reference structure. This model system is used as a testing ground for the evolutionary selection. For each evolutionarily selected sequence, 10 simulated forward-folding runs are tried with different random-number implementation. Then, out of the 10 structures thus obtained, 3 structures which have the smallest rms deviations from the reference structure are picked up. The averaged rms deviation of these 3 structures from the reference structure is used as a measure of the folding ability of the sequence. When the average rms value is $< 5 \text{ \AA}$, the criterion used to select the sequence would be based on sensible design rules.

Random Walk in the Sequence Space

The result of threading the 3icb sequence and other sequences in the library onto the 3icb reference structure is shown in the histogram of Fig. 3. The 3icb sequence has the lowest energy and other structurally unrelated sequences look like “random sequences” with the energy distribution width δE .

This is further confirmed by actually generating random sequences. Starting from the native sequence (original 3icb sequence), a position in the sequence is randomly selected and replaced with the arbitrarily chosen type of residue. This “point mutation” is iterated many times. This random walk trajectory quickly goes into the energy region where other library sequences are located and wanders around there. We may conclude that when threaded onto the certain structure, most of the 20^N possible sequences have random energy with the width δE . Some sequences which occupy a tiny portion of the vast sequence space may be located at the low-energy edge of this random distribution. The number of these sequences,

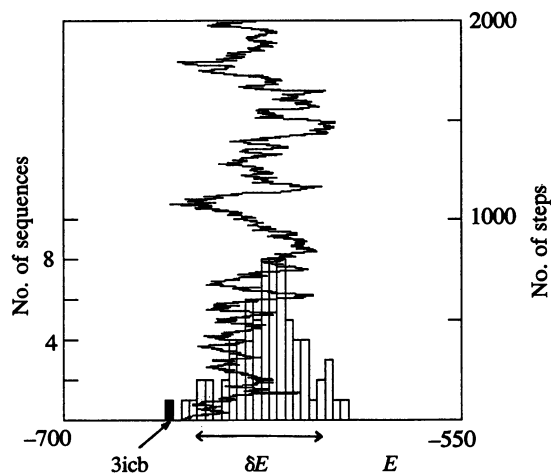


FIG. 3. Results of threading sequences in the library onto the 3icb reference structure are shown by the histogram. The energy fluctuation along the sequence-space random walk is superposed.

however, could still be extremely large. Then how many such sequences exist? How is the energy landscape in the sequence space? To get insights into these questions, random sequences are selected with several selection criteria in the next section.

Inverse Folding with Evolutionary Selection

The cost function H is used to select sequences. A random point mutation is generated and the change in H , δH , is evaluated. The chain conformation is fixed in the reference structure. When $\delta H \leq 0$, this mutation is accepted, but when $\delta H > 0$, the mutation is rejected with the probability $1 - \exp(-\delta H/T)$ and is accepted otherwise. This is a Metropolis Monte Carlo algorithm with the “selection temperature” T . Sequences are selected with different criteria by using different cost functions. Starting from a random sequence, the sequence which minimizes H is found by gradually lowering T . Such simulated annealing-selection runs are examined below with five different selection criteria.

Criterion I. The cost function is energy, $H = E$. The energy of the selected sequence quickly becomes lower than the energy of the native sequence and $> 80\%$ of the residues are replaced by Trp. This is so because the Trp–Trp distance is small in most cases found in the library and the Trp–Trp potential has a deep minimum at a short distance (6 \AA for the class 9 potential).

Criterion II. To exclude the frequent Trp pair appearance, the cost function is set to be $H = |E - E^{\text{native}}|$, where E^{native} is the energy of the reference structure threaded with the native 3icb sequence. After the cost function is lowered enough, there still exists large fluctuation in sequence; the cost-function landscape is almost flat around $H = 0$. With these sequences, however, the reference structure is unstable; with application of forces derived from Eq. 4, the reference structure is deformed to other irrelevant structures. Thus the naive energy minimization in the sequence space does not yield a meaningful result.

Criterion III. To assure the stability of the reference structure, the following cost function is considered:

$$H = |E - E^{\text{native}}| + a \left[\sum_{i=1}^N (\partial E / \partial \mathbf{r}_i)^2 \right]^{1/2} + b \left[\sum_{\alpha=1}^{20} (\omega_{\alpha} - \omega_{\alpha}^{\text{native}})^2 \right]^{1/2}, \quad [5]$$

where the second term is proportional to the force strength at the reference structure. ω_{α} and $\omega_{\alpha}^{\text{native}}$ with $\alpha = 1-3N$ are eigenvalues obtained by diagonalizing the matrices $\partial^2 E / \partial \mathbf{r}_i \partial \mathbf{r}_j$ and $\partial^2 E^{\text{native}} / \partial \mathbf{r}_i \partial \mathbf{r}_j$, respectively. Since the higher-lying eigenvalues depend on the details of the potential and should not be important in the sequence comparison, only the 20 low-lying eigenvalues are used. It is harder to make the second and third terms small than to make the first term small; therefore, a relatively large value, $a = b = 10$, is used.

With this cost function, the evolutionary trajectory does not reach the sequence with $H = 0$ but travels among the local minima with low H values; the cost-function landscape is rugged around $H = 0$. The fluctuation of sequence identity between the generated sequences and the native 3icb sequence is shown in Fig. 4a. To examine the folding ability, some of the generated sequences are picked up and for each of them the forward-folding test is performed with Brownian-motion MD. The average rms values obtained with this test are plotted in Fig. 4b. The average rms deviation is $> 10 \text{ \AA}$ for most of the sequences; the chain is easily frozen to the metastable deformed structure having distorted helices. Thus the stability against small conformational fluctuations at the optimized

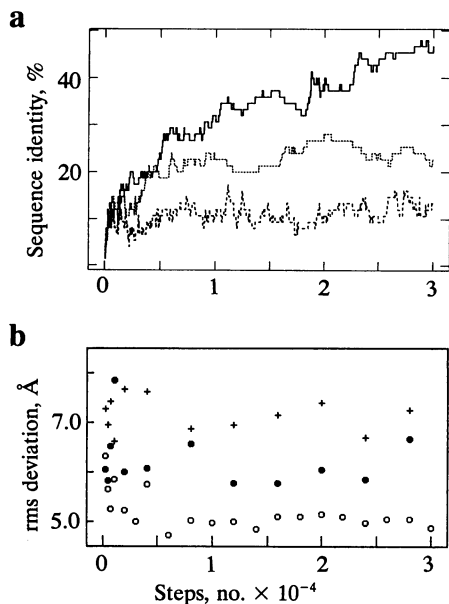


FIG. 4. (a) Fluctuation of the sequence identity of selected sequences to the native 3icb sequence. The solid line shows the sequences selected with criterion V, the dotted line shows those selected with criterion IV, and the dashed line shows those selected with criterion III. (b) The average rms deviation from the reference structure is shown as a measure of the folding ability of selected sequences. ○, Sequences selected with criterion V; ●, sequences selected with criterion IV; and +, sequences selected with criterion III.

structure is not enough to assure the kinetic preference of the structure. Therefore, not only the small fluctuations but also the large-amplitude deformation of the structure must be taken into account.

Criteria IV and V. To take account of distant conformations, the following cost function is considered:

$$H = \sum_{\mu=1}^{10} P(\mu) \left\{ |E_{\mu} - E_{\mu}^{\text{native}}| + a \left[\sum_{i=1}^N (\partial E_{\mu} / \partial \mathbf{r}_i - \partial E_{\mu}^{\text{native}} / \partial \mathbf{r}_i)^2 \right]^{1/2} + b \left[\sum_{\alpha=1}^{20} (\omega_{\alpha} - \omega_{\alpha}^{\text{native}})^2 \right]^{1/2} \right\} \quad [6]$$

where 10 different conformations $\mu = 1-10$ are used; $\mu = 10$ is the reference structure, and the conformations $\mu = 1-9$ are structures 1-9 in Fig. 2. E_{μ} and E_{μ}^{native} are energies at the μ th conformation with the tested and the native sequences, respectively. $P(\mu)$ is $P(\mu) = \exp(-cE_{\mu}/T) / \sum_{\mu} \exp(-cE_{\mu}/T)$, and c is chosen to be a rather small value, $c = 0.01$ (criterion IV) or 0.001 (criterion V), so that multiple conformations are efficiently sampled even at the low selection temperature. The sequence-selection calculations are started from a random sequence with the selection temperature $T = 0.1$. At this initial temperature, in criterion IV, $P(1) = 0.036$, $P(2) = 0.047$, $P(3) = 0.060$, $P(4) = 0.111$, $P(5) = 0.121$, and $P(6) = 0.121$, and in criterion V, $P(1) = 0.091$, $P(2) = 0.094$, $P(3) = 0.096$, $P(4) = 0.102$, $P(5) = 0.103$, and $P(6) = 0.103$. Thus the stretched conformations have greater importance in criterion V.

With criterion IV, H stays around 25% of the initial value after 5000 mutation steps; the cost-function landscape is rugged as is the case with criterion III. With criterion V,

however, H continues to become smaller during the simulation. Thus, in the case of the criterion V, there seems to be a fairly deep minimum of the cost function though the trajectory has not yet reached that minimum during the simulation. Examples of sequences generated with criterion V are shown in Fig. 5. They share two important features with the native sequence: (i) Positions of glycine are the same in the simulated and the native sequences, except around residues 57-59, and (ii) there is a close similarity in the hydrophobicity pattern between the simulated and the native sequences. Though sequences fluctuate much along the evolutionary trajectory, these two features remain unchanged. With criterion III or IV, however, the pattern similarity is less evident, and with criterion II the pattern matches poorly with that of the native sequence. By changing parameters in Eqs 3 and 4, the parameter dependence of the results can be tested. Within the range of reasonable parameter values, however, the qualitative results shown in Fig. 4 are not altered.

The results of the forward-folding test are shown in Fig. 4b. Criterion IV is not sufficient to select the foldable sequences. The resulting conformations often include the bent helices and the average rms deviation is $>5 \text{ \AA}$. With criterion V, on the other hand, the generated sequences have folding ability as high as that of the native sequence. Along the evolutionary trajectory, the average rms deviation rapidly decreases, and at around 5000 mutation steps it becomes $<5 \text{ \AA}$. It should be noted that at around 5000 steps the sequence identity to the native sequence is still as low as 25%. It is interesting that the minimum sequence identity between two homologous conformations found in the Protein Data Bank is about 25-30% for $N \approx 60-80$ (31). Thus the evolutionary trajectory first finds the foldable sequences at this "homology threshold" area and then proceeds to the sequence space which is closer to the native sequence.

Comparing five different selections, we find that not only the local but also the global features in the conformation space have to be considered. It is especially important to take the energy landscape of the extended conformations into consideration (Fig. 6). When the sequence is selected so as to have an energy surface similar to the native one along the folding

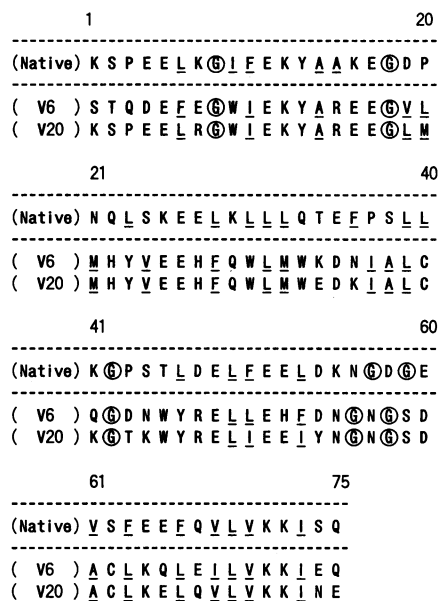


FIG. 5. Selected sequences compared with the native 3icb sequence. V6 is the sequence generated at 6000 mutation steps in the evolutionary trajectory selected with criterion V, and V20 is the sequence generated at 20,000 mutation steps with criterion V. Glycines are marked with circles and hydrophobic residues are underlined.

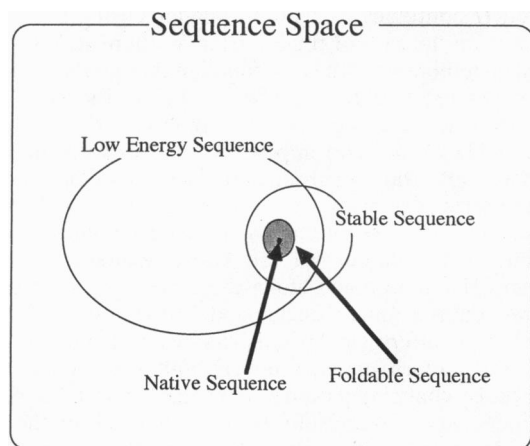


FIG. 6. Schematic representation of the amino acid sequence space. There are many low-energy sequences which are unrelated to the target structure. The foldable sequences belong to a subgroup of both the low-energy sequences and the sequences which can stabilize the target structure against small fluctuations.

pathway, the chain can fold with as high probability as the native chain.

With criterion V a number of foldable sequences are easily found that have 25–50% sequence identity to the native sequence. Preliminary analysis suggests that the speed of the sequence alternation is not uniform along the trajectory. Statistical properties of these sequences need to be investigated to see how the neutral drift among sequences is possible.

The forward folding kinetics of the present system are characterized by hierarchical structural ordering due to the coexistence of different length scales (helices, turns, HLH motifs, and stacking of two HLH motifs). To reach the correct conformation, structures in various length scales have to develop and competitions arise among them. The chain visits many conformations until this structural conflict is minimized. This competition and cooperation are observed with a wide range of parameter values and are important features of the present model system. When the trajectory fails to fold, some structural order often develops too rapidly, without waiting for the growth of structures in other length scales. These results are in accord with the earlier observation of Go (32), who stressed the importance of consistency among the different length scales. The energy surface which supports such hierarchical structural ordering should also have a hierarchical nature; the energy surface should lead to the “funnel” along which the structural conflicts among orders in different length scales are avoided. The present inverse-folding results suggest that sequences must be selected so as to reproduce this hierarchical nature of the energy surface. Especially the comparison between selection criteria IV and V suggests that the sequence should be designed so that the primordial forms of various structural orders are prepared before the collapse to the globule state.

The method is still limited to the helix-bundle proteins. There is much room, however, to improve the Hamiltonian we used here. The dihedral-angle restriction should be taken into account and forces proportional to the surface area should be used. The context-dependent forces are also important (29, 30). The present simulated-evolution experiments have shown that the sequence design is closely related to the design of the

hierarchical energy surface. Thus, improvement of folding algorithms, finding of new design principles, and understanding of the diversity of protein structures will progress in a synergistic way by analyzing the global structures of the energy landscape. There, the combined forward- and inverse-folding simulations with the sophisticated knowledge-based Hamiltonian will play an important role.

I gratefully acknowledge the helpful comments of T. Noguti and M. Go. This work was partially supported by a grant-in-aid from the Ministry of Education, Science, and Culture, Japan.

1. Levinthal, C. (1968) *J. Chim. Phys.* **65**, 44–45.
2. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.
3. Bryngelson, J. D. & Wolynes, P. G. (1989) *J. Phys. Chem.* **93**, 6902–6915.
4. Sasai, M. & Wolynes, P. G. (1990) *Phys. Rev. Lett.* **65**, 2740–2743.
5. Sasai, M. & Wolynes, P. G. (1992) *Phys. Rev. A* **46**, 7979–7997.
6. Garel, T. & Orland, H. (1988) *Europhys. Lett.* **6**, 307–310.
7. Shakhnovich, E. I. & Gutin, A. M. (1989) *Europhys. Lett.* **8**, 327–332.
8. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 8721–8725.
9. Socci, N. D. & Onuchic, J. N. (1994) *J. Chem. Phys.* **101**, 1519–1528.
10. Camacho, C. J. & Thirumalai, D. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 6369–6372.
11. Miller, R., Danko, C. A., Faselka, M. J., Balazs, A. C., Chan, H. S. & Dill, K. A. (1992) *J. Chem. Phys.* **96**, 768–780.
12. Chan, H. S. & Dill, K. A. (1993) *J. Chem. Phys.* **99**, 2116–2127.
13. Sali, A., Shakhnovich, E. I. & Karplus, M. (1994) *Nature (London)* **369**, 248–251.
14. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72**, 3907–3910.
15. Fraunfelder, H. & Wolynes, P. G. (1994) *Phys. Today* **47** (2), 58–64.
16. Bowie, J. U. & Eisenberg, D. (1993) *Curr. Opin. Struct. Biol.* **3**, 437–444.
17. Lau, K. F. & Dill, K. A. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 638–642.
18. Yue, K. & Dill, K. A. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4163–4167.
19. Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 1942–1946.
20. Shakhnovich, E. I. & Gutin, A. M. (1991) *J. Theor. Biol.* **149**, 537–546.
21. Shakhnovich, E. I. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 7195–7199.
22. Wodak, S. J. & Rooman, M. J. (1993) *Curr. Opin. Struct. Biol.* **3**, 247–259.
23. Sippl, M. J. (1992) *J. Mol. Biol.* **213**, 859–883.
24. Hendlich, M., Lackner, P., Weitkusch, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1992) *J. Mol. Biol.* **216**, 167–180.
25. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Nature (London)* **358**, 86–89.
26. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–170.
27. Luthy, R., Bowie, J. U. & Eisenberg, D. (1992) *Nature (London)* **356**, 83–85.
28. Wilmanns, M. & Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 1379–1393.
29. Friedrichs, M. S., Goldstein, R. A. & Wolynes, P. G. (1991) *J. Mol. Biol.* **222**, 1013–1034.
30. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
31. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
32. Go, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.