Proteins with Selected Sequences Fold into Unique Native Conformation

E. I. Shakhnovich
Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138
(Received 1 December 1993)

We design sequences of 80-monomer model protein which provide very low energy in the target ("native") structure. Then the designed sequence is subjected to lattice Monte Carlo simulation of folding. In all runs model protein folded from random coil to the unique native conformation, effectively "solving" the multiple minima problem. These results suggest that thermodynamically oriented selection of sequences which makes the native conformation a pronounced deep minimum of energy solves the problem of kinetic accessibility of this conformation as well.

PACS numbers: 87.15.Da, 61.43.–j, 64.60.Cn, 64.60.Kw

The complexity of the protein folding problem is in the fact (often referred to as Levinthal paradox [1]) that unique, native conformation should be chosen in the folding process without scanning the astronomic number of possible conformations. The important question is whether this kinetic ability of natural proteins to fold is due to evolutionary selection of their sequences and, if yes, how can this feature be encoded in a protein sequence?

The straightforward approach (tried, e.g., in [2]) would be to take natural amino-acid sequence and simulate a (simplified) model of a protein expecting convergence to the native 3D conformation. However, the difficulty with this approach is that protein sequences could have been evolutionary designed to fold to their native structures with some "exact" set of potentials while simulations necessarily use approximate energetics [3] for which the native structure may be neither a global nor a pronounced local minimum. It is then hard to expect any folding algorithm to converge to a "native" state which may not be distinguished by energy from many other conformations.

This suggests the idea of using protein design to study folding of model one-domain proteins of realistic size. The goal is to design a sequence which has very low energy in a given (arbitrary) conformation. Folding simulation with the same potential function as was used at the design stage will then reveal whether this conformation can be reached in a reasonable time. Combination of design and folding "in one pair of hands" makes it possible to address the basic questions of protein folding and evolution separately from the problem of finding the correct potential functions for protein simulations.

In this study we model proteins as positioned on a cubic lattice. The Hamiltonian of a model protein is determined by the set coordinates of its monomers $\{r_i\}$ and (quenched) sequence of monomers $\{\sigma_i\}$ which denotes the identity of each monomer. Contact approximation is taken for the Hamiltonian,

$$E(\{\sigma_i\},\{r_i\}) = \frac{1}{2}\sum_{i,j}^{N} U(\sigma_i\sigma_j)\Delta(r_i - r_j), \qquad (1)$$

where N is the total number of monomers and Δ defines the contact potential between them: $\Delta(r) = 1$ if monomers are lattice neighbors and 0, otherwise. We consider our model proteins positioned on a cubic lattice with unit bond length.

The set of potentials $U(\alpha,\beta)$ characterizes energies with which a monomer of type α interacts with a monomer of type β. First we tried two-letter sequences (hydrophobic-hydrophilic) like ones used in two-dimensional lattice models of proteins [4]. However, two-letter sequences appeared to be inappropriate for studying protein folding in three-dimensional models (see below). Therefore in what follows twenty-letter representation of protein sequences was used. In this case $U(\alpha,\beta)$ is a 20×20 matrix; as an example we used the one derived by Miyazawa and Jernigan [3] from the statistical distribution of contacts in native proteins.

In this work we studied folding of 80-monomer chains. Following the idea to combine folding and design we choose (arbitrarily) a target structure which is in our case a compact conformation of a chain on the cubic lattice. An example of the target structure is shown in Fig. 1.

After the target structure is picked, sequence design should be made to find a sequence which fits the target structure with low energy as determined according to Eq. (1) where coordinates $\{r_i\}$ correspond to target conformation. To this end the sequence-space Monte Carlo (MC) procedure of design was used [5,6]. The idea of sequence design is very simple: For the design purposes just view Eq. (1) as one where coordinates of the target structure $\{r_i\}$ are quenched but sequence variables $\{\sigma_i\}$ are annealed and Eq. (1) should be optimized with respect to them. This leads naturally to the idea of simulated annealing in sequence space; the procedure is straightforward and the details are published in [5,6].

The following argument based on the theory of heteropolymers allows us to estimate whether the native (target) structure corresponds to the global energy minimum for the designed sequence.

We divide the set of all conformations into two groups: the ones which have significant similarity with the tar-

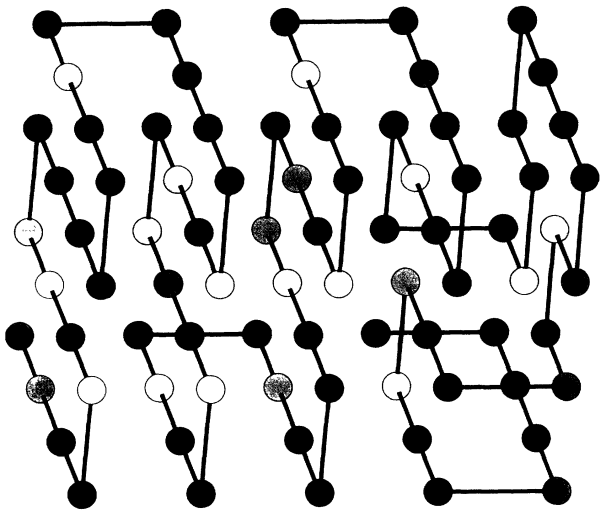


FIG. 1. An example of a compact conformation of an 80-monomer on a cubic lattice and the optimized sequence. Amino acids of different types are shown by different gray scale for illustrative purposes. This conformation as well as several other conformations with their sequences (not shown) were used as native structures in our studies. The shown sequence was designed to have low energy in the shown conformation.

get structure and the remaining vast majority of conformations which have marginal or no similarity with the target structure, just like two randomly superimposed conformations.

For conformations which are not similar to the target structure the designed sequence is effectively random and therefore the statistics of their energies are equivalent to those of a random heteropolymer. (A similar argument was first given by Bryngelson and Wolynes in their discussion of the "minimal frustration" model of protein folding [7].)

The important feature of random heteropolymers is that there exists a threshold energy E_c such that the probability to find conformations with energy well below E_c is extremely small [8–11]. Therefore the successful design should create sequences whose energy E_N in the native conformation is well below E_c: In this case random conformations (structurally nonsimilar to the native state) will not have energies close to that of the native conformation and therefore will not serve as deep energetic traps for folding.

E_N is known directly from Eq. (1) for the designed sequence. To estimate E_c we use the replica mean-field theory of heteropolymers [8–11]. $E_c = E_0 - JN(2\ln\gamma)^{1/2}$, where γ is the number of conformations per monomer. The important parameters E_0 and J are the mean and the standard deviation of interaction energies. Since we are using parameters which are obtained from protein statistics, we have only relative energies and do not know the absolute energy scale for those parameters. So we use the energy unit at which $J = 1$. This requires multiplication of all parameters by a scaling factor. To determine this scaling factor we generated a set of 1000 random sequences (all having the same amino-acid composition) and fitted them into the target structure adjusting the scaling factor so that $J^2 = (\langle E^2\rangle - \langle E\rangle^2)/N = 1$. $\langle E\rangle = E_0$ and $\langle\,\rangle$ denotes averaging over the set of random sequences. We took $\gamma = 3.5$ which takes into account excluded volume and certain degree of compactness of unfolded conformations for which variance of interactions J is estimated. The estimates were done for two sets of parameters: "two-letter" code with monomers of two types ("H" and "P") so that $U(H,H) = -1$; $U(H,P) = U(P,P) = 0$ and the twenty-letter set of Miyazawa and Jernigan. The amino-acid composition was set to be 50% H and 50% P monomers for two-letter chains and corresponding to averaged composition in proteins [12] for the twenty-letter set. The results for 80-monomer chains are given below.

(1) Two-letter heteropolymers: $E_c = -72.3$, $E_N = -61$. The model is not specific enough to have unique structure: All possible energy levels are multiple degenerate. No folding to unique structure is possible in that case.

(2) Twenty-letter parameters: $E_c = -123.6$, $E_N = -156.5$. The estimated gap is pronounced, $\approx 23T$ at the temperature at which most of the simulations have been done. In all that follows twenty types of monomers are used and the results are reported for that model.

Now we simulate folding of the designed sequence using the simple lattice Monte Carlo folding algorithm [13–17] and energy function given by Eq. (1). The move set which we used allows corner flips and crankshaft motions but excludes multiple occupancies of lattice sites. It was argued in [17] that such a move set makes cubic lattice simulation ergodic.

Simulations started from random coil conformations. There was made a total of 1000 runs starting from different randomly chosen coil conformations. The main result of this work is that in each run chain folded into the unique target conformation with mean first passage folding time close to 10^6 MC steps.

A typical folding trajectory recorded at temperature $T = 1$ is shown in Fig. 2. Analysis of energy changes with Monte Carlo time shows that structures with energy lower than the energy of the native state have never been encountered. In order to estimate whether this conclusion is sensitive to the move set we repeated simulation with enhanced move set which allowed also for 3,4,5-monomer crankshaft moves. The results are similar: Again the target state was the lowest energy one, and the trajectory was similar to the one shown in Fig. 2.

Pronounced fluctuations around the minimum energy structure make this model close to the Molten Globule (MG) [18,19]. This is due to the fact that there are

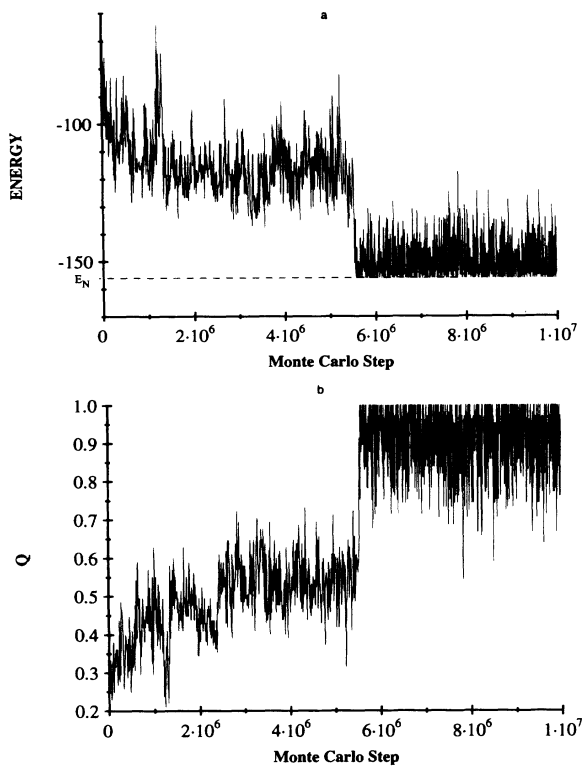


FIG. 2. A typical MC trajectory of folding simulations for 80-monomer chain. (a) The dependence of energy on MC step. The energy of the native conformation is shown as E_N. (b) The dependence of normalized number of native contacts on MC step. The maximal number of contacts $N_{\text{total}} = 105$ for a compact 80-monomer. For each conformation we normalize the number of native contacts, Q, by N_{total} so that $Q = 1$ corresponds to the native conformation.

no side chains in the model, which tight packing distinguishes the MG from the native state and makes the native conformation more rigid [19].

The ability to fold appears to be a virtue of designed sequences and is temperature dependent, as expected. Steepness of the curves in Fig. 3 is consistent with the assertion [5] that designed sequences have a first-order folding transition. Applied to proteins this suggests that the coil-MG transition may be also first order, like the native-MG one. The first-order character of the native-MG transition, however, may be due to a different reason, side-chain freezing [19], which is not considered in the present model (see [20] for the discussion of first-order transitions in macromolecules).

The temperature dependence of entropy can be obtained from temperature dependence of energy $E(T)$ using the thermodynamic relation

$$s(T) = s(\infty) + \frac{1}{N}\left(\frac{E(T)}{T} - \int_T^\infty \frac{E(t)}{t^2}dt\right). \quad (2)$$

Here $s(\infty)$ is a high-temperature (athermal) limit of entropy. The value $s(\infty) = \ln(4.68) + \frac{1}{6}\ln(79)/79$ is known since at high T it coincides with that for an athermal polymer on a cubic lattice [21]. Our simulations were

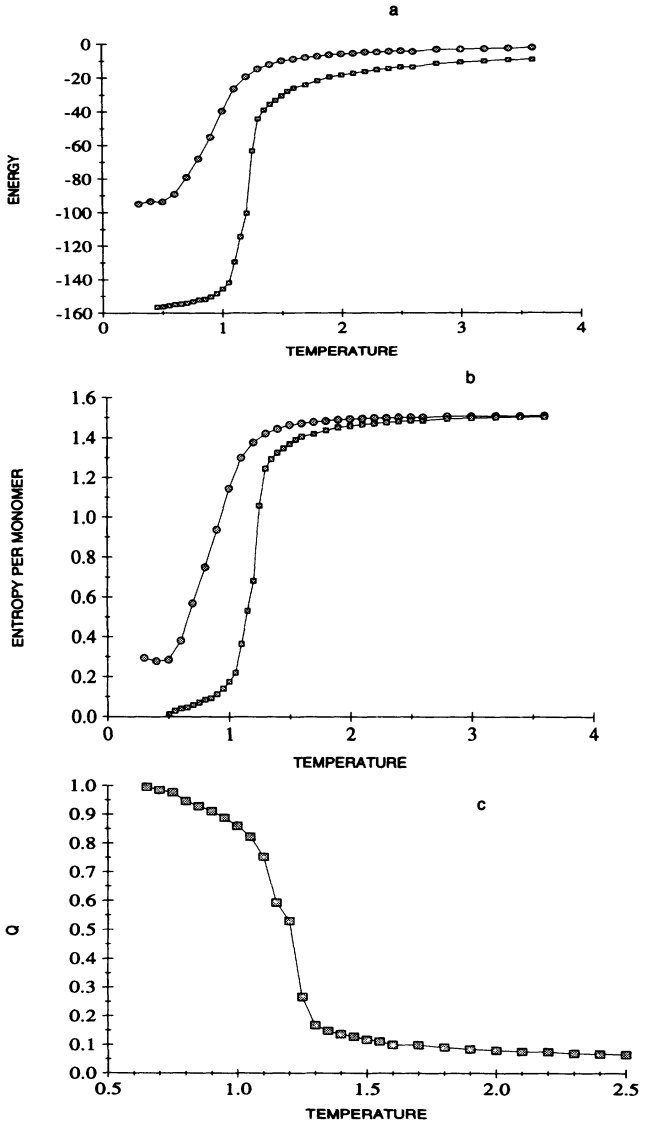


FIG. 3. Temperature dependence of energy E (a), configurational entropy per monomer (b), and structural similarity with the native state (c) for the designed sequence (squares) and for a random sequence with the same amino-acid composition as the designed one (circles). At each temperature 10^8 MC steps were made, and average energy E and structural similarity with the target conformation Q were determined as an average over the whole run at a given temperature. The calculation of the entropy curve is explained in the text.

performed in the temperature range $0.5 < T < 10.0$. We took $s(T = 10) = s(\infty)$. Only part of the temperature dependence corresponding to the temperature range $0.5 < T < 3.6$ is shown to provide a reasonable scale to show the transition. The truncated part at $T > 3.6$ is a trivial base line. In the low-temperature limit $s(T = 0.5) = 0.007$. The smallness of this number is consistent with the main result of this work—that designed sequences repetitively return to the target (native) conformation.

The same procedure was used then to calculate confor-

mational entropy of the random sequence. The number of conformations is determined as usually $M = \exp(Ns)$. In this case the same rate of annealing leads to freezing without development of unique structure: Different runs end up in different, unrelated conformations. The number of such frozen low-temperature conformations is estimated from the calculation of entropy (Fig. 3) to be $\sim 10^9$. Note also that even in the denatured state energy of designed sequence is noticeably lower than that of the random sequence.

Low-temperature freezing for a random sequence is a kinetic phenomenon: It was shown in [22] that in this case the global minimum cannot be reached by *any* algorithm in a reasonable [less than "Levinthal" $\exp(\alpha N)$] time. This does not contradict the assertion [9] that random sequences can have a thermodynamically stable unique structure in a certain temperature range. The reason is that the unique structure of random sequences becomes thermodynamically stable only at temperatures lower than T_c, the glass transition temperature [7–11,22,23]. However, as was shown in [22] (see also the excellent discussion in [23]), at $T < T_c$ the kinetics become extremely slow because the ruggedness of energy landscape of random sequences turns out to be crucial at temperatures lower than T_c. Sequences with large gaps have native structures which are stable at $T > T_c$, resolving therefore the contradiction between the requirement of thermodynamic stability and kinetic accessibility which is characteristic of random sequences.

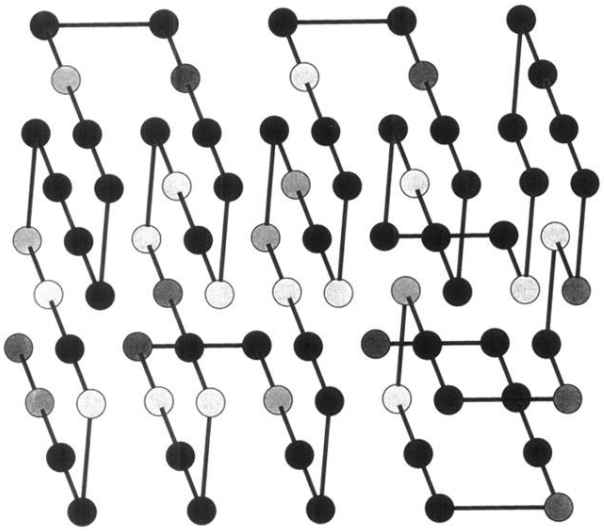
Analysis of the curve $Q(T)$ in Fig. 3(c) suggests that the native state is sufficiently stable at temperatures at which simulations were done. For example, at $T = 0.8$, $Q \approx 0.95$ which means that 95% of native contacts persist throughout the simulations. Conformations which have 95% of native contacts differ from the native one (shown, e.g., in Fig. 1) by "tails" of 3–4 monomers long stretching out of the native structure. The alternative interpretation of this result would be that the chain spends 95% of the time in the native state and 5% of the time in unfolded conformation. The analysis of simulation data at $T = 0.8$ suggests that the chains spend practically all the time in or near native conformation, so that short-tail fluctuations account for the fact that $Q < 1$. This can be also illustrated from the estimate of entropy at $T = 0.8$, $S = 0.05$ per monomer, which suggests that fluctuations cover $\sim$ 100 conformations, each only slightly (by 3–4 monomers) different from the native state. This is consistent with the "short-tail stretching" picture.

The same experiments were repeated with several other sequences and several other randomly chosen target structures for proteins of different lengths (36–100 monomers). One target conformation even had a quasi-knot (Abkevich, Grosberg, and Shakhnovich, unpublished results). In all cases the results of simulations are qualitatively the same and are quantitatively close to the ones presented in this work.

Our design procedure generated sequences for which the target structure is likely to be the global (or at least accessible stable local) energy minimum separated by a pronounced energy gap from the set of non-native conformations. It is remarkable to note that such thermodynamically oriented design solved at the same time the *kinetic* problem making the native structure also kinetically accessible. This may represent a simple and universal principle of evolutionary selection of one-domain proteins with stable and kinetically accessible native conformation.

I am grateful to Victor Abkevich, Alexander Gutin, Martin Karplus, Peter Leopold, Oleg Ptitsyn, and Andrej Sali for interesting discussions. Graphic program ASGL by Andrej Sali was used to generate some of the plots. This work was supported by the Packard Foundation.

[1] C. Levinthal, J. Chem. Phys. **65**, 44 (1968).
[2] C. Wilson and S. Doniach, Proteins: Struct. Funct. Genetics **6**, 193 (1989).
[3] S. Myazawa and R. Jernigan, Macromolecules **18**, 534 (1985).
[4] K.F. Lau and K.A. Dill, Macromolecules **22**, 3986–3997 (1990).
[5] E.I. Shakhnovich and A.M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[6] E.I. Shakhnovich and A.M. Gutin, Protein Eng. **6**, 793 (1993).
[7] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[8] E.I. Shakhnovich and A.M. Gutin, Biophys. Chem. **34**, 187 (1989).
[9] E.I. Shakhnovich and A.M. Gutin, Nature (London), **346**, 773 (1990).
[10] C. Sfatos, A. Gutin, and E. Shakhnovich, Phys. Rev. E **48**, 465 (1993).
[11] A. Gutin and E. Shakhnovich, J. Chem. Phys. **98**, 8174 (1993).
[12] T. Creighton, *Proteins Structure and Molecular Properties* (Freeman, San Francisco, 1992).
[13] A. Sali, E.I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[14] E.I. Shakhnovich, G.M. Farztdinov, A.M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[15] R. Miller, C. Danko, M.J. Fasolka, A.C. Balazs, H.S. Chan, and K.A. Dill, J. Chem. Phys. **96**, 768 (1992).
[16] C. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369–6372 (1993).
[17] H.J. Hilhorst and J.M. Deutch, J. Chem. Phys. **63**, 5153 (1975).
[18] O.B. Ptitsyn, in *Protein Folding* (Freeman, New York, 1992), Chap. 6, pp. 243–300.
[19] E.I. Shakhnovich and A.V. Finkelstein, Biopolymers **28**, 1667 (1989).
[20] M. Karplus and E. Shakhnovich, in *Protein Folding* (Ref. [18]), Chap. 4, p. 127.
[21] P.G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, NY, 1970).
[22] J.D. Bryngelson and P.G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[23] H. Fraunefelder and P.G. Wolynes, Phys. Today **47**, 58 (1994).



NDLVGIRYATLTVIYAAQLAGIRAHDLGVVKNPEVVFI
KYTELGTIRATDLSFFKAPVHSVASFSCGLCMLERLVC

FIG. 1. An example of a compact conformation of an 80-monomer on a cubic lattice and the optimized sequence. Amino acids of different types are shown by different gray scale for illustrative purposes. This conformation as well as several other conformations with their sequences (not shown) were used as native structures in our studies. The shown sequence was designed to have low energy in the shown conformation.