How does a protein fold?

Andrej Šali, Eugene Shakhnovich & Martin Karplus*

Department of Chemistry, 12 Oxford Street,
Harvard University, Cambridge, Massachusetts 02138, USA
* To whom correspondence should be addressed

THE number of all possible conformations of a polypeptide chain is too large to be sampled exhaustively. Nevertheless, protein sequences do fold into unique native states in seconds (the Levinthal paradox). To determine how the Levinthal paradox is resolved, we use a lattice Monte Carlo model in which the global minimum (native state) is known. The necessary and sufficient condition for folding in this model is that the native state be a pronounced global minimum on the potential surface. This guarantees thermodynamic stability of the native state at a temperature where the chain does not get trapped in local minima. Folding starts by a rapid collapse from a random-coil state to a random semi-compact globule. It then proceeds by a slow, rate-determining search through the semi-compact states to find a transition state from which the chain folds rapidly to the native state. The elements of the folding mechanism that lead to the resolution of the Levinthal paradox are the reduced number of conformations that need to be searched in the semi-compact globule ($\sim 10^{10}$ versus $\sim 10^{16}$ for the random coil) and the existence of many ($\sim 10^{3}$) transition states. The results have evolutionary implications and suggest principles for the folding of real proteins.

The mechanism of protein folding is not understood, despite many studies devoted to this subject[1]. The essential question is how a polypeptide chain is able to fold rapidly, in milliseconds to seconds, to the native state, despite the very large number of conformations that exist for the chain (Levinthal paradox)[2]. To approach this problem, we used a simplified model that consists of a 27-bead self-avoiding chain on a cubic lattice[3] (Fig. 1). The native (lowest energy) state can be determined exactly[3] and a survey of the folding behaviour of many sequences is possible[4]. In addition, the full phase space density of the system can be obtained and the thermodynamic properties can be calculated as a function of the folding 'reaction coordinate'. The model is sufficiently complex that a resolution of the Levinthal paradox is required for folding: some sequences find the native state in only $\sim 10^{7}$ Monte Carlo steps even though there are $\sim 10^{16}$ possible conformations (Fig. 3*b*). Because the lattice model does not include the amino-acid side chains, the process considered here may correspond to the folding of real proteins to the molten globule state; that is, the molten globule, if it has a defined fold, is the native state in the present model.

In the first part of the analysis, 200 sequences with random interactions were generated and subjected to Monte Carlo folding simulations (Fig. 2)[4]. Of these, 30 chains found the known native state in a short time. These chains correspond to actual protein sequences in the sense that they fold on a reasonable timescale; the remaining sequences do not represent protein sequences because they do not fold rapidly enough, although they do have a unique minimum. These sequences serve as controls. The 30 folding sequences were analysed and compared with the non-folding sequences. Several suggested mechanisms for resolving the Levinthal paradox do not apply to the present model as the features assumed to be responsible for rapid folding are found to be the same for the folding and non-folding sequences. These include a high number of short- versus long-

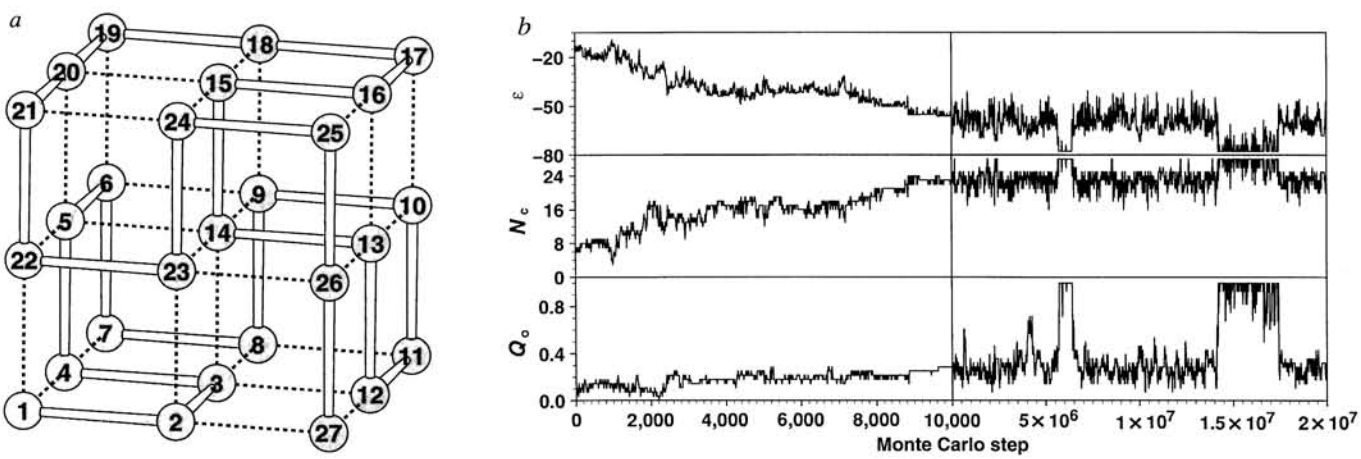

FIG. 1 Lattice model of protein folding. *a*, An example of a compact self-avoiding structure of a chain of 27 monomers (filled numbered circles) with 28 contacts (dashed lines). The structure shown is the native state of the folding random sequence 43 used as an example in this paper. The total energy of a conformation is the sum of contact energies: $E=\sum_{i<j}\Delta(r_i, r_j)B_{ij}$, where r_i are the positions of monomers i, B_{ij} are the contact energies for pairs of monomers i, j, and $\Delta(r_i, r_j)$ is 1 if monomers i and j are in contact and is 0 otherwise; two monomers are in contact if they are not successive in sequence and at unit distance from each other. The values of the B_{ij} are obtained from a gaussian distribution with a mean B_0 and standard deviation σ_B. This particular model for B_{ij} corresponds to a heteropolymer with a random sequence of monomers of many different types whose heterogeneity is measured by σ_B[11]. The parameter B_0 is an overall attractive term that emulates the hydrophobic effect observed in globular proteins. A particular sequence is defined by the matrix of contact energies B_{ij}. The B_{ij}s correspond to the contact energies in real proteins[4], such as those described by Miyazawa and Jernigan[19]. The native conformation is the compact self-avoiding chain with the lowest energy. *b*, Typical trajectory for a folding random sequence ($T=1.3$). Energy, ε (in units of k_BT where k_B is the Boltzmann constant); the number of contacts, N_c; fraction of the number of contacts in common with the native state, Q_0. The instantaneous values of these quantities are plotted every 10 Monte Carlo (MC) steps in the first part of the trajectory ($\leqslant$10,000 MC steps) and every 20,000 steps in the subsequent part. The folding trajectory starts with a random-coil conformation and consists of local MC moves of one or two successive monomers that preserve bond lengths and avoid multiple occupancy of the lattice sites[4]; one MC step corresponds to a move and a test of its acceptance with the Metropolis criterion[20]. In the first part of the trajectory, ~50% of the MC moves are accepted, whereas only 5–10% of the moves are successful in the subsequent part of the simulation.

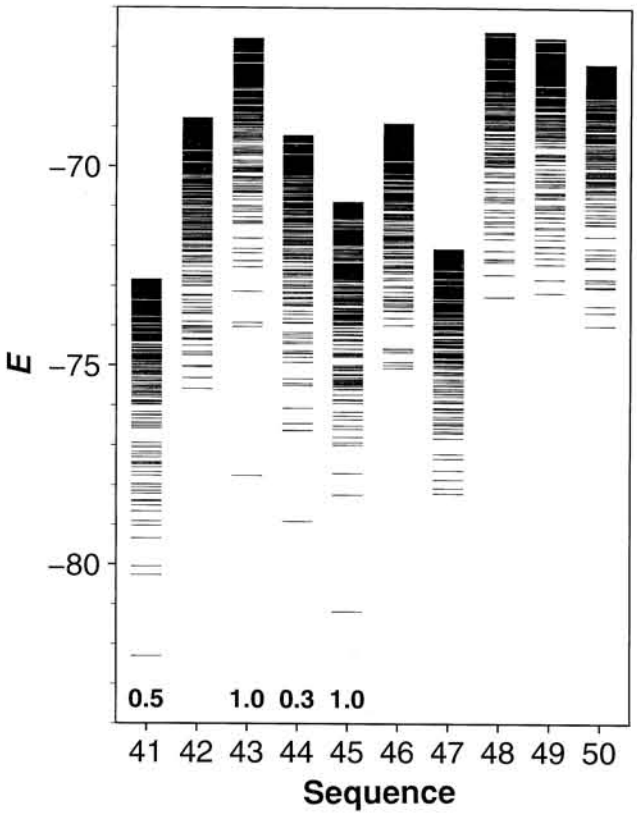

FIG. 2 Energy spectra for 10 folding and non-folding random sequences. The energies of the 400 lowest compact self-avoiding conformations are shown. The native state corresponds to the bottom bar. The numbers below the spectra show the folding tendencies of the corresponding sequences; if no number is given, folding tendency is 0. A sequence folds in a given MC simulation if it finds the native conformation within 50×10^6 MC steps. Folding tendency of a given sequence is defined as the fraction of 10 MC runs that started with a random conformation and reached the native conformation under a given set of conditions. A sequence is a folding sequence if the native conformation is structurally unique and folding tendency is high ($\geqslant 0.4$) under conditions where the native structure is thermodynamically stable. A sequence is a non-folding sequence if the folding tendency is 0.0. There are 24 intermediate sequences that we do not consider here. Optimal values for parameters $B_0(-2)$ and $\sigma_B(1)$ were determined by exhaustive sampling of folding tendency as a function of these two parameters[4]. Each sequence is studied at a temperature where the native state has a high probability to be reasonably thermodynamically stable; the native state has a weight larger than 0.2 relative to other compact self-avoiding conformations[4].

range contacts in the native state[6], a high content of secondary structure in the native state[7], a strong correlation between the native contact map and the interaction parameters[8], and the existence of a high number of low-energy states with near-native conformations[3]. Moreover, there is no repetitive trapping of the non-folding sequences in the same local minimum, so that the native state cannot be a metastable state[9]. The only significant difference between folding and non-folding sequences is that the native state is at a pronounced energy minimum (Fig. 2). This is the necessary and sufficient condition for a sequence to fold rapidly in the present model.

To explore the mechanism of folding, the density of states was determined as a function of the energy and the reaction coordinate (the number of native contacts) (Fig. 3); a related reaction coordinate, the number of residues in the native conformation, has been used in other models of folding[10]. From the density of states, the mean energy, entropy and free energy are calculated (Fig. 4*a*–*c*). Above the critical temperature ($T=1$), the energy and entropy decrease smoothly as the chain approaches the native state. The free energy, by contrast, has a maximum corresponding to the transition region that separates the denatured and native states. This barrier, which makes the reaction a cooperative transition, is dominated by the entropic contribution, as can be seen from the temperature dependence of the free energy. Below the critical temperature, the free-energy reaction profile becomes rugged[10,11] and folding is very slow because the chain is likely to be trapped in one of the many local minima.

The time history of the folding process (Fig. 1*b*) shows that there is a rapid collapse in about 10^4 Monte Carlo steps to a semi-compact random globule; the number of contacts increases, the energy decreases, whereas the fraction of native contacts remains below 0.3. In this way, the total of $\sim 10^{16}$ random-coil conformations is reduced to $\sim 10^{10}$ random semi-compact globule states. The fast collapse results from a large energy gradient (Fig. 1*b*) and the presence of many empty lattice points. In the second stage (Fig. 4*d*), which is rate-limiting, the chain searches for one of the $\sim 10^3$ transition states. The transition region consists of all states from which the chain folds rapidly to the native state. The transition states are structurally similar to the native state, with 23–26 (80–95%) of the native contacts. Generally, the mean first passage time, τ, for finding any of the n states among the total of N states by a random search which explores r states per unit of time is $N/(n \times r)$. For the present model, $\tau \sim 10^{10}/(10^3 \times 1) = 10^7$ (Fig. 3*c*), identical to the observed timescale. This indicates that the rate-limiting stage in folding consists of a random search for a transition state in the semi-compact part of the phase space; a folding 'pathway' is not involved in finding the native state. In the third stage, the chain rapidly (within $\sim 10^5$ Monte Carlo steps) attains the native conformation from any one of the transition states.

The pronounced energy minimum is the necessary condition for the folding of a 27-mer on a lattice because it guarantees that the native state is stable above the critical temperature, where the rearrangements with local unfolding required in the rate-limiting stage are energetically possible. It is also a sufficient condition because a random search of compact globules with random structures can rapidly find a transition state that folds to the stable native state in a short time. Although a non-folding sequence may also fold slowly to its native state above the critical temperature, it would not be stable at that temperature and therefore could not correspond to a real protein.

The size of the search for the native state is greatly reduced when the chain is semi-compact, as it is in real proteins[12]. Nevertheless, in the 27-mer model, as in real proteins, a random search of a collapsed globule cannot find the native state in the observed time. It is the existence of a transition region, consisting of a large number of states, that reduces the search time to realistic values. Extrapolation of the folding time to real proteins suggests that the three-stage random-search mechanism could be effective

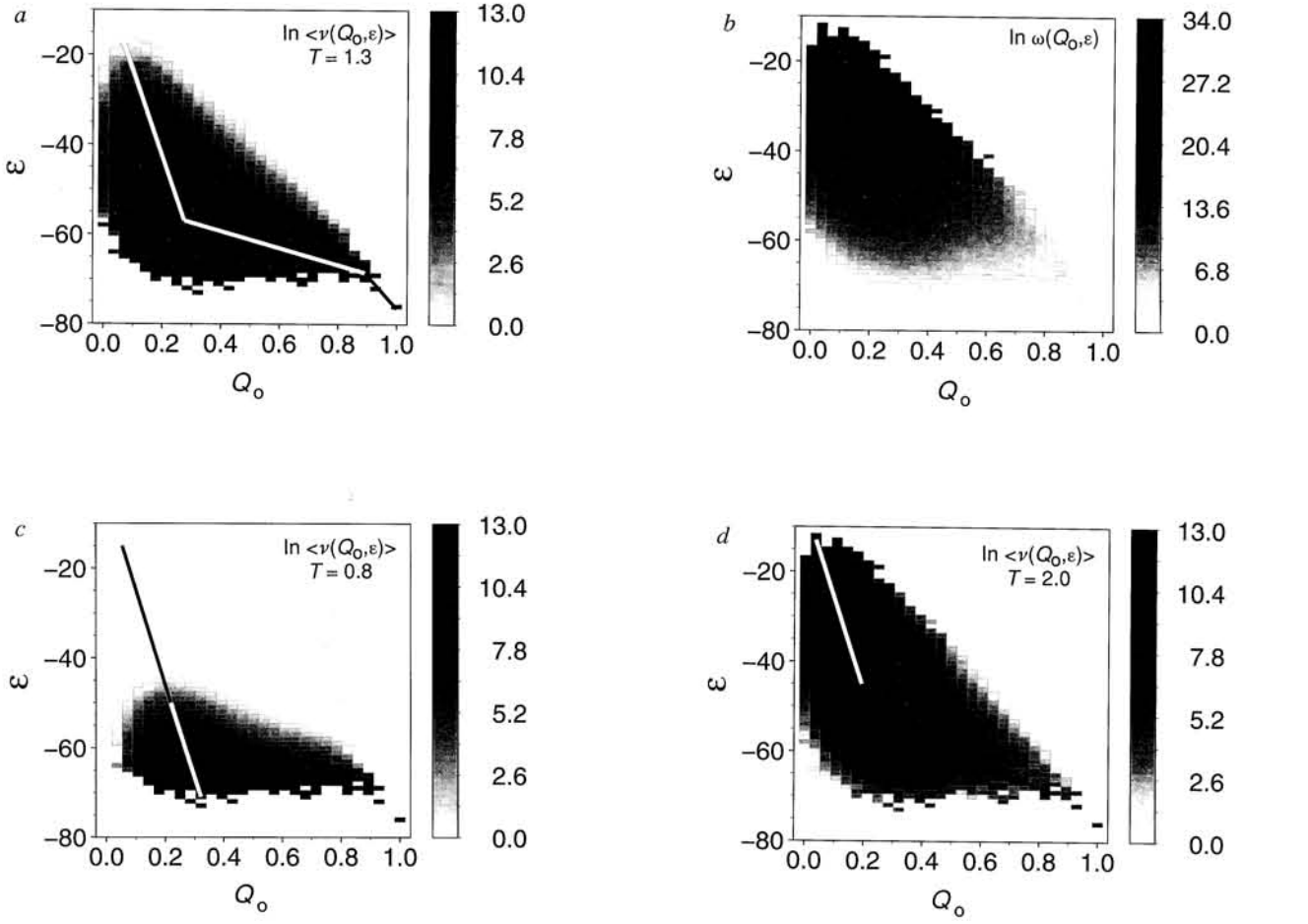

FIG. 3 Density of states for a folding random sequence, as obtained from a long Monte Carlo simulation. *a*, The logarithm of the average occupancy of the (Q_o, ε) bins, $\langle \nu(Q_o, \varepsilon) \rangle$, at $T=1.3$. Each bin spans $1/28$ Q_o units and 1 energy unit. The occupancy of a bin is the number of Monte Carlo steps that remain or result in any of the conformations corresponding to the bin. The average is calculated from 20 independent MC sampling simulations of 50×10^6 steps each (A.Š, E.S. and M.K., manuscript in preparation). The broken line indicates a typical folding pathway. *b*, The logarithm of the density of states, $\omega(Q_o, \varepsilon)$. The density of states is calculated from $\langle \nu(Q_o, \varepsilon) \rangle$ by the use of $\langle \nu(Q_o, \varepsilon) \rangle / \langle \nu(Q_o=1, \varepsilon=\varepsilon_o) \rangle = \omega(Q_o, \varepsilon)/\omega(Q_o=1, \varepsilon=\varepsilon_o) \exp[-\varepsilon/(k_BT)]/\exp[-\varepsilon_o/(k_BT)]$ where ε_o is the energy of the native state and k_B is the Boltzmann constant set to 1; because $\omega(Q_o=1, \varepsilon=\varepsilon_o)=1$, the density of states is $\omega(Q_o, \varepsilon)=\langle \nu(Q_o, \varepsilon) \rangle / \langle \nu(1, \varepsilon_o) \rangle \exp(-(\varepsilon_o-\varepsilon)/(k_BT))$. The calculation is accurate because thermodynamic equilibrium at the present Q_o, ε resolution is achieved, as demonstrated by a small fractional error in $\langle \nu(Q_o, \varepsilon) \rangle$ (less than 10% in the populated parts of the phase space). The summation of $\omega(Q_o, \varepsilon)$ over all Q_o and ε results in about 1.1×10^{16} self-avoiding states; this is in good agreement with the extrapolation of the exact enumerations for shorter chains: $\nu^{L-1}/12 = 2.2 \times 10^{16}$ where L is the number of monomers, $\nu=4.68$ is the average number of monomer states, and division by 12 corrects for symmetry[21]. Semi-compact states are defined as the states with energy less than $-45k_BT$; summation over the appropriate region of $\omega(Q_o, \varepsilon)$ yields $\sim 10^{10}$ such states. The transition states correspond to all the states with $0.8 \leqslant Q_o < 1$; there are $\sim 10^3$ such states. *c*, The average occupancy of the bins at a low $T(T=0.8)$ as calculated from $\omega(Q_o, \varepsilon)$. The line shows the rapid collapse of a random coil chain into a random semi-compact frozen conformation. *d*, The calculated average occupancy of the bins at a high $T(T=2.0)$. The line shows a rapid collapse of the random coil chain into a region consisting of many interconverting semi-compact random states; the ground state is no longer stable so the chain spends most of its time in the entropically favoured denatured states. In contrast, at $T=1.0$, the native state occurs for ~40% of the time while still being kinetically accessible.

for small proteins (Fig. 4*d*). However, the mechanism breaks down for long chains because the folding time increases exponentially with chain length owing to the fact that the number of semi-compact states increases faster than the number of the transition states. Thus, a modification of the present mechanism is required for larger proteins.

It is possible that proteins existing early in evolution were small enough to fold according to the three-stage random-search mechanism[13]. Because the pre-biotic and early biotic environment was hot, unusually thermostable proteins were required, such as those found in the most primitive bacteria that live at temperatures at high as 105 °C[14]. If so, the stability condition required a native state that corresponded to a deep energy minimum for which the folding problem was solved simultaneously according to the present mechanism. As evolution progressed, longer proteins evolved. These proteins had to fold on the same timescale. One way of achieving this is by evolving proteins with sequences that have a larger difference between the native and non-native contact energies than the random folding sequences of the present model. Such sequences would have an even more pronounced global minimum in their potential surfaces, in line with the 'consistency principle'[15] and the 'principle of minimal frustration'[16]. It is also in accord with the existence of a nucleus for folding[6], or the early appearance of secondary structural elements[7,17], neither of which are found in the present model. As a result, the collapse would not result in random semi-compact

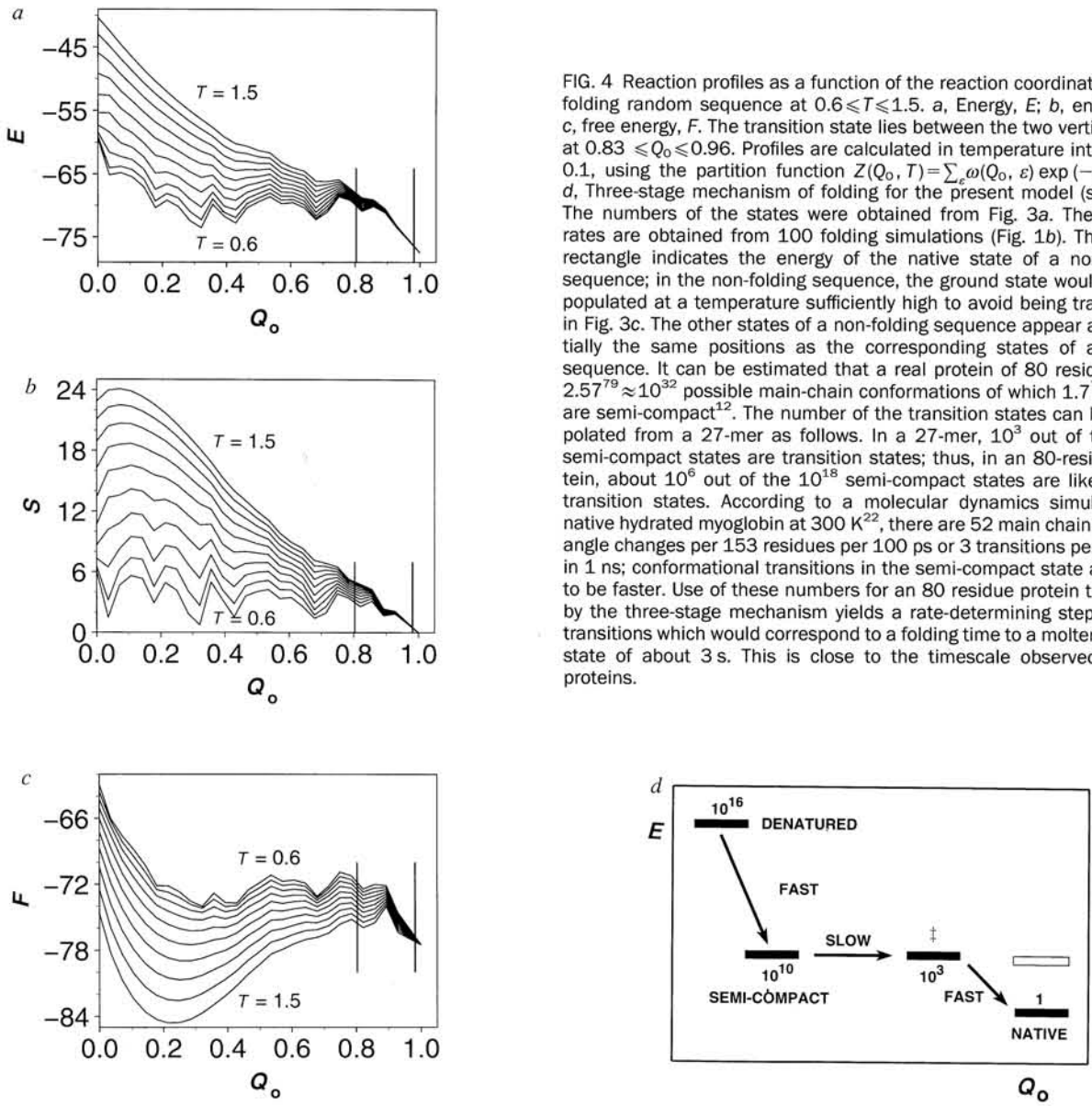

FIG. 4 Reaction profiles as a function of the reaction coordinate for the folding random sequence at $0.6 \leqslant T \leqslant 1.5$. *a*, Energy, *E*; *b*, entropy, *S*; *c*, free energy, *F*. The transition state lies between the two vertical lines at $0.83 \leqslant Q_0 \leqslant 0.96$. Profiles are calculated in temperature intervals of 0.1, using the partition function $Z(Q_0, T) = \sum_\varepsilon \omega(Q_0, \varepsilon) \exp(-\varepsilon/(k_B T))$. *d*, Three-stage mechanism of folding for the present model (see text). The numbers of the states were obtained from Fig. 3*a*. The relative rates are obtained from 100 folding simulations (Fig. 1*b*). The empty rectangle indicates the energy of the native state of a non-folding sequence; in the non-folding sequence, the ground state would not be populated at a temperature sufficiently high to avoid being trapped as in Fig. 3*c*. The other states of a non-folding sequence appear at essentially the same positions as the corresponding states of a folding sequence. It can be estimated that a real protein of 80 residues has $2.57^{79} \approx 10^{32}$ possible main-chain conformations of which $1.7^{79} \approx 10^{18}$ are semi-compact[12]. The number of the transition states can be extrapolated from a 27-mer as follows. In a 27-mer, 10^3 out of the 10^{10} semi-compact states are transition states; thus, in an 80-residue protein, about 10^6 out of the 10^{18} semi-compact states are likely to be transition states. According to a molecular dynamics simulation of native hydrated myoglobin at 300 K[22], there are 52 main chain dihedral angle changes per 153 residues per 100 ps or 3 transitions per residue in 1 ns; conformational transitions in the semi-compact state are likely to be faster. Use of these numbers for an 80 residue protein that folds by the three-stage mechanism yields a rate-determining step of 10^{12} transitions which would correspond to a folding time to a molten globule state of about 3 s. This is close to the timescale observed in real proteins.

globules and the very favourable native contacts would lead more directly to a native-like molten globule state; the folding pathway in Fig. 3*a* would approximate a straight line from the random-coil to the 'native state'. Such a mechanism has been observed in folding simulations of long chains with highly stable native states[18] (V. Abkevich, A. Gutin and E.S., unpublished results; A. Dinner, A.Š. and M.K., unpublished results). Actual proteins could use an intermediate mechanism that might vary with the external conditions.

The results of this study may have implications for the prediction of the structure of a protein from its amino-acid sequence. The success of the three-stage random search in finding the pronounced global minimum on a potential surface suggests, at least for small proteins, that the bottleneck in structure prediction may be the derivation of a suitable potential function rather than the design of folding algorithms. □

Received 6 January; accepted 24 March 1994.

1. Creighton, T. E. (ed.) *Protein Folding* (Freeman, New York, 1992).
2. Levinthal, C. in *Mossbauer Spectroscopy in Biological Systems* (eds Debrunner, P., Tsibris, J. C. M. & Münck, E.) 22–24 (Univ. Illinois Press, Urbana, 1969).
3. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. *Phys. Rev. Lett.* **67,** 1665–1668 (1991).
4. Šali, A., Shakhnovich, E. I. & Karplus, M. *J. molec. Biol.* **235,** 1614–1636 (1994).
5. Harding, M., Williams, D. & Woolfson, D. *Biochemistry* **30,** 3120–3128 (1991).
6. Wetlaufer, D. B. *Proc. natn. Acad. Sci. U.S.A.* **70,** 697–701 (1973).
7. Kim, P. & Baldwin, R. *A. Rev. Biochem.* **59,** 631–660 (1990).
8. Gō, N. & Abe, H. *Biopolymers* **20,** 1013–1031 (1981).
9. Honeycutt, J. D. & Thirumalai, D. *Biopolymers* **32,** 695–709 (1992).
10. Bryngelson, J. D. & Wolynes, P. G. *J. phys. Chem.* **93,** 6902–6915 (1989).
11. Shakhnovich, E. I. & Gutin, A. M. *Biophys. Chem.* **34,** 187–199 (1989).
12. Dill, K. A. *Biochemistry* **24,** 1501–1509 (1985).
13. Di Iorio, E. E. *et al. Proc. natn. Acad. Sci. U.S.A.* **90,** 2025–2029 (1993).
14. Stetter, K. O. in *Frontiers of Life* (eds Trân Thanh Vân, J. K., Mounolou, J. C., Schnieder, J. & McKay, C.) 195–212 (Editions Frontières, Gif-sur-Yvette, France, 1992).
15. Gō, N. & Abe, H. *Adv. Biophys.* **18,** 149–164 (1984).
16. Bryngelson, J. D. & Wolynes, P. G. *Proc. natn. Acad. Sci. U.S.A.* **84,** 7524–7528 (1987).
17. Karplus, M. & Weaver, D. L. *Nature* **260,** 404–406 (1976).
18. Taketomi, H. & Gō, N. *Int. J. Peptide Prot. Res.* **7,** 445–449 (1975).
19. Miyazawa, S. & Jernigan, R. L. *Macromolecules* **18,** 534–552 (1985).
20. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. *J. chem. Phys.* **21,** 1098–1092 (1953).
21. Sykes, M. F. *J. chem. Phys.* **39,** 410–412 (1963).
22. Loncharich, R. J. & Brooks, B. R. *J. molec. Biol.* **215,** 439–455 (1990).

ACKNOWLEDGEMENTS. We thank A. Dinner, D. Šali, O. Ptitsyn, A. Gutin, G. Archontis, A. Caflisch, O. Becker and L. Demetrius for discussions concerning the protein folding problem. A.Š. is a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research. This investigation was aided by a grant from The Jane Coffin Childs Memorial Fund for Medical Research (A.Š.), by David and Lucille Packard Fellowship (E.S.), and by a grant from the National Science Foundation and a gift from Molecular Simulations Inc. (M.K.). The computations were done on IBM RS/6000, Silicon Graphics Iris 4D, SUN Sparcstation, DEC Decstation, DEC Alphastation, and NeXT workstations.