Protein Folding Studied by Monte Carlo Simulations

Andrej Sali*, Eugene Shakhnovich and Martin Karplus [1]

Dept. of Chemistry, 12 Oxford St, Harvard University, Cambridge, MA 02138, U.S.A.
* Present address: The Rockefeller University, 1230 York Avenue, New York, NY 10021, U.S.A.

Abstract

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 (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 close to the native 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 ($\approx 10^{10}$ *versus* $\approx 10^{16}$ for the random coil) and the existence of many ($\approx 10^{3}$) transition states. The results have evolutionary implications and suggest principles for the folding of real proteins.

[1] to whom correspondence should be addressed. Phone: +1 (617) 495 4018, Fax: +1 (617) 496 3204, E-mail: marci@tammy.harvard.edu

1 Introduction

The mechanism of protein folding is not understood, despite many studies devoted to this subject [1, 2, 3, 4]. The essential question is how a polypeptide chain is able to fold rapidly, in ms to s, to the stable native state in spite of the very large number of possible conformations that exist for the chain (Levinthal paradox) [5]. The mechanisms of protein folding are proposed on the basis of conceptual discussions [6, 7, 8, 9], simulations [10, 11, 12, 13, 14], and theory [15, 16, 17]. Theories of protein folding were recently reviewed in references [1, 2, 3, 4, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28].

While the dynamics of the native state can be characterized relatively well by molecular dynamics simulations [29], much less is known about the potential surface governing the non-native portion of conformation space that is involved in protein folding. It includes a wide range of structures which may differ by tens of angstroms. Concomitantly, instead of the time scale of picoseconds to nanoseconds that is required for exploring the neighborhood of the native state, the characteristic times corresponding to the motions in the full conformation space are in the nanosecond to second range. The existence of such a separation of time and length scales with fast local motions and slow large-scale motions makes it possible to introduce two simplifying concepts, which can serve as a basis for theoretical work on protein folding. The first simplification is an effective potential or potential of mean force and the second is a discretized description of the polypeptide chain. Both of these concepts are based on the idea of "preaveraging" the small-scale motions to obtain a "coarse grained" model, which can treat a molecule on the time and length scales at which protein folding occurs. This leads to simplified models of proteins that include only a subset of atoms [30] and to discretized conformational space of various lattice models [31, 32, 33] that employ Monte Carlo (MC) dynamics to simulate the kinetics.

Recently, lattice models have been used to address a variety of aspects of the protein folding problem [9, 10, 12, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45]. In one particularly simple class of lattice models, the protein chain is represented as a string of beads on the 2D square lattice [33, 46] or 3D cubic lattice [16]. These models are generally not meant to simulate folding of any particular real amino acid sequence. However, the lattice models do capture the most essential features of proteins, including their heteropolymeric nature and the interactions between sequentially local and distant amino acid residues. Thus, the simple lattice models may be suitable for elucidating overall features of the folding process such as the main stages of folding. The advantages of simple lattice models are that simulations of many sequences are possible and that frequently the global energy minimum can be determined by enumeration.

As the model does not include side chains, the "native" state in the lattice model corresponds to a compact globule with the native fold of a real protein. Such structured globules may correspond to the experimentally observed molten globules, which (although expanded relative to the native state) preserve much of the backbone structure, and whose side chains are free to undergo dihedral angle transitions [47, 48, 49, 50]. Molten globules appear to be a late stage in the folding of some proteins and their formation involves resolution of the Levinthal paradox [48, 50, 51]. For further discussion of the suitability of lattice models to describe real proteins and of MC to represent molecular dynamics, see refs. [4, 38, 52, 53].

In this review, we summarize our results obtained by the use of MC simulations of 27-mer heteropolymers on a cubic lattice. We identified the features of the folding sequences that allow them to fold rapidly into a unique and stable native state [52] and the mechanism by which they achieve this state [12, 54].

[1] Abbreviations used: MC, Monte Carlo; 3SRS-mechanism, three-stage random search mechanism.

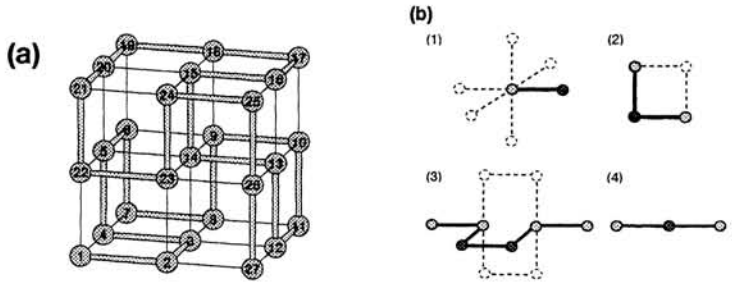

Figure 1: Lattice model of protein folding. *a*, An example of a compact self-avoiding structure of a chain of 27 monomers (filled numbered circles) with 28 contacts (thin lines). The structure shown is the native state of the folding random sequence 43 used as an example in this paper. The total energy 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 are at unit distance from each other. The values of the B_{ij} are obtained from a Gaussian distribution with a mean B_o 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 [15]. The parameter B_o 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}. In terms of their magnitude and standard deviation, the B_{ij}'s correspond to the contact energies in real proteins [52], such as those described by Miyazawa and Jernigan[72]. The *native* conformation is the compact self-avoiding chain with the lowest energy. *b*, The three types of possible MC moves (1–3). Situation 4 shows a conformation where no move of the central monomer is possible. The current conformation is shown in thick lines. Possible new conformations are shown in dashed lines. A move is possible if all new positions are unoccupied. The monomers that are being moved are shown in dark gray.

2 Methods

We use a simplified model that consists of a 27-bead self-avoiding chain on a cubic lattice [10] (Figs. 1 and 2). The native (lowest-energy) state can be determined exactly [10] and a survey of the folding behavior of many sequences is possible [52]. In addition, the full phase space density of the system (Fig. 6) [12, 54] can be obtained and the thermodynamic properties can be calculated as a function of the folding "reaction coordinate" (Fig. 7) [12, 54]. The model is sufficiently complex that the Levinthal paradox exists: *i.e.*, some sequences find the native state in only $\approx 10^7$ MC (MC) steps even though there are $\approx 10^{16}$ possible conformations (Fig. 6b).

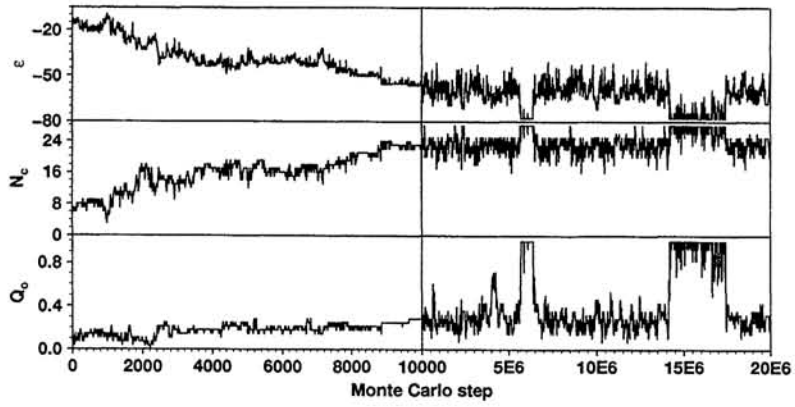

Figure 2: Typical trajectory for a folding random sequence ($T = 1.3$). Energy, ϵ (in units of k_BT); the number of contacts, N_c; fraction of the number of contacts in common with the native state, Q_o. The instantaneous values of these quantities are plotted every 10 MC steps in the first part of the trajectory ($\leq$ 10000 MC steps) and every 20000 steps in the subsequent part. The folding trajectory starts with a random-coil conformation and consists of local MC moves described in Fig. 1b[52]; one MC step corresponds to a move and a test of its acceptance with the Metropolis criterion[73]. In the first part of the trajectory, $\approx$ 50% of the MC moves are accepted, while only 5–10% of the moves are successful in the subsequent part of the simulation.

3 Results and Discussion

3.1 A pronounced energy minimum is necessary and sufficient for rapid folding to a stable native state

In the first part of the analysis, 200 sequences with random interactions were generated and subjected to MC folding simulations (Fig. 3) [52]. Of these, 30 chains found the known native state in a short time. These chains correspond to actual protein sequences in the present model; the remaining sequences, which do not fold, do not correspond to protein sequences and serve as controls. The 30 folding sequences were analyzed and compared with the non-folding sequences. Several suggested mechanisms for resolving the Levinthal paradox do not apply to the present model; *i.e.*, the features assumed to be responsible for rapid folding in these mechanisms are found to be the same for the folding and non-folding sequences. These include a high number of short *versus* long range contacts in the native state [8], a high content of secondary structure in the native state [7], a strong correlation between the native contact map and the interaction parameters [31], and the existence of a high number of low energy states with near-native conformations [10]. 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 [55]. The only significant difference between the folding and non-folding sequences is that the native state is at a pronounced energy minimum (Fig. 3). This is the necessary and sufficient condition for a sequence to fold rapidly in the present model (Fig. 4). It is necessary because no sequences without a pronounced energy minimum fold to the native state and it is sufficient because all sequences with such a minimum do fold.

An essential element of the study was to examine a large number of sequences and separate

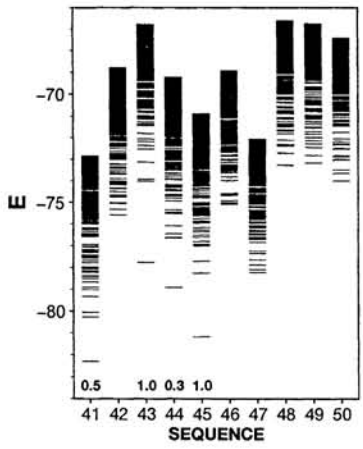

Figure 3: Energy spectra for 10 folding and non-folding random sequences. The energies of the 400 lowest compact self-avoiding conformations are shown. The native state corresponds to the bottom bar. The numbers below the spectra show the foldicities of the corresponding sequences; if no number is given, foldicity is 0. A sequence *folds* in a given MC simulation if it finds the native conformation within $50 \cdot 10^6$ MC steps. *Foldicity* 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 foldicity is high (≥ 0.4) under conditions where the native structure is thermodynamically stable. A sequence is a *non-folding sequence* if the foldicity is 0.0. There are 24 intermediate sequences that we do not consider here. Optimal values for parameters B_o (-2) and σ_B (1) were determined by exhaustive sampling of foldicity as a function of these two parameters[52]. Each sequence is studied at a temperature, T_x, slightly above the midpoint of the folding transition[52]. An order parameter that describes a transition of a chain from a degenerate state with many backbone conformations to a state with few, possibly only one, backbone conformation is $X(T) = 1 - \sum_i^M p_i^2$, where $p_i = \exp(-\frac{\epsilon_i}{k_B T})/Z$ and $Z = \sum_i^M \exp(-\frac{\epsilon_i}{k_B T})$. k_B is the Boltzmann constant set to 1 in this study, p_i is the Boltzmann probability for a system to be in state i, M is the number of all compact self-avoiding states, and Z is the configuration partition function of the chain. T_x is defined such that $X(T_x) = 0.8$ where the native state has a weight almost invariably larger than 0.2 relative to other compact-self avoiding conformations.

those that fold from those that do not. Consequently, a temperature at the neighbourhood of T_m, the mid-point of the folding transition, was used to speed up the reaction and save computer time (Fig. 3). Optimization of the folding rate was important to verify that certain sequences did not fold. If folding to the native state were always possible under the simulation conditions, there would not have been any nonfolding sequences and our computer experiment would have failed. Figure 5 shows the temperature dependence of the folding rate over a range where 0 to 50% of the polymers are in the native state at equilibrium. Linear folding kinetics is observed throughout the temperature range; this justifies the use of a relatively high temperature for the folding experiments (Fig. 3).

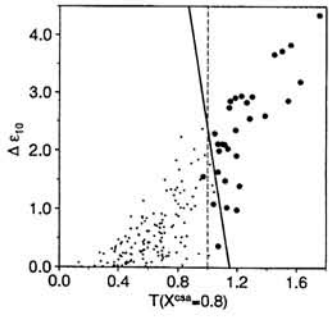

Figure 4: Discrimination between strongly folding (circles) and non-folding sequences (dots). $\Delta\epsilon_{10}$ is the energy gap between the native state and the second most stable compact self-avoiding conformation. T_x is the temperature at which the MC simulations are performed. The larger are $\Delta\epsilon_{10}$ and T_x, the more pronounced the global energy minimum is. The continuous line separates the two groups of sequences by minimizing the number of the folding and non-folding sequences in the non-folding and folding parts of the plot, respectively. The dashed line indicates the critical temperature, defined as the temperature below which almost no folding is observed [52]. T_x is determined for each sequence separately such that the native state had a high probability to be reasonably stable (see Fig. 3) [52].

The reason for the correlation between folding and stability is that significant portions of the potential energy surface of the model system are "rugged". In particular, the random collapsed state that is sampled in the three-stage random search (3SRS) mechanism (see below) is a multiminimum surface on which the search for the native state requires surmounting many intervening barriers. This can be done on a reasonable time scale only if the folding temperature is sufficiently high for there to be a significant probability of overcoming such barriers. However, at a high temperature, the majority of the random sequences have a ground state that is not stable; in other words, the Boltzmann probability of being in any of the excited states is too large unless a sizeable energy gap separates the native state from the excited states. As the temperature at which the folding simulations are done is near the midpoint of the thermodynamic transition temperature between the native and denatured states, the simulation temperature is high enough to overcome the barriers only for the strongly folding sequences with a particularly stable ground state . The transition temperature for the nonfolding sequences is so low that the 27-mer gets trapped in a metastable well. This qualitative argument only explains why the pronounced energy minimum is necessary for folding. The explanation for why it is also sufficient is provided by the 3SRS mechanism discussed below. It is likely that the necessity of the energy gap condition is general for reasonable lattice models and for real proteins, while its sufficiency may be of more limited applicability.

The importance of the temperature in the protein folding reaction and the relation between an energy gap and folding, which are clearly demonstrated by the 27-mer simulations [12, 52], have been discussed previously. Based on statistical mechanical arguments and spin glass theory, Bryngelson and Wolynes [17, 56] suggested that there are two temperatures that need to be considered in determining the folding properties of a sequence. One is the folding

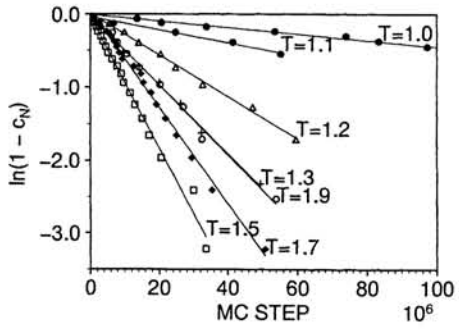

Figure 5: Kinetics of conversion of the denatured state into the native state for sequence 43 as a function of temperature. Kinetics of conversion of the denatured state into the native state for sequence 43 as a function of temperature. For each temperature, the probability, C_N, that the native state is reached in a given number of MC steps is estimated from 100 folding trials. The lines are linear least-squares fits to the points determined at the temperature as indicated on the plot. It can be seen that the results satisfy the simple unimolecular rate equation $\mathrm{d}C_N/\,\mathrm{d}t = kC_D$, where k is the rate constant, C_D is the probability of being in the denatured state. The line at $T = 1.0$ is linear in the whole time range explored (up to $400 \cdot 10^6$ MC steps where $C_N \approx 1$).

temperature and the other is the glass transition temperature. They suggested that the glass transition temperature corresponds to the temperature below which the chain is frozen into a random low energy conformation because it does not have enough energy to overcome the barriers separating such conformations. On this basis, they concluded that the temperature at which the sequence folds must be higher than the glass transition temperature. Further, it was shown that random sequences do not satisfy this condition and so would be likely to be trapped in metastable states [17, 56]. Bryngelson and Wolynes introduced specific biases toward the native state to make folding possible. The existence of such biases on the entire potential surface corresponds to the principle of "minimum frustration" [17, 56], which is closely related to the "consistency" or "harmony" principle proposed by Gõ and Abe [57]. One way of introducing the bias is by the use of associative-memory Hamiltonians [58, 59, 60, 61, 62] which have been employed successfully in a variety of applications; e.g., it was shown that the ratio between the folding and glass transition temperatures, whose maximization was assumed to lead to faster folding, is proportional to the ratio of the energetic separation of the native state from the denatured states and the range of energies corresponding to the denatured states [58].

Simulations using other simple lattice models confirmed that the pronounced global energy minimum is associated with rapid folding [38, 40, 63]. An additional confirmation came from the success of designing stable folding sequences by minimizing the energy of a given native state [35, 63, 64, 65].

3.2 Random sequences fold by the three-stage random search mechanism

The insights concerning the role of the temperature and the energy gap provided by the 27-mer computer experiment [52] and theoretical analyses [17] do not, in themselves, provide the specific mechanism by which a model resolves the Levinthal paradox. For this purpose, further examination of the results for the strongly folding sequences in the 27-mer was required [12, 54].

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. 6); a related reaction coordinate, the number of residues in the native conformation, has been used previously [17]. If there were a nucleus of a small number of native contacts that would lead rapidly and with high probability to the native state, the fraction of native contacts would not be a good reaction coordinate; instead, the fraction of the contacts present in the nucleus should be used. However, a folding nucleus is not present in the model, confirming that the fraction of native contacts is a suitable reaction coordinate. From the density of states, the mean energy, entropy, and free energy for a given temperature were calculated (Fig. 7). Above the critical temperature ($T_c \approx 1$), the energy and entropy decrease smoothly as the chain approaches the native state. The free energy 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 reaction profile corresponding to the free energy becomes rugged [15, 17] and folding is much slower because the chain is likely to be trapped in one of the many local minima.

Analyses of individual trajectories for both strongly folding and non-folding sequences, as well as the calculated density of states and reaction profiles [12, 54], demonstrated that the sequences in the present model fold according to the three-stage random search (3SRS) mechanism at temperatures above 1.0 (Fig. 8). The time history of the folding process (Fig. 1b) shows that there is a rapid collapse in $\approx 10^4$ MC steps to a semi-compact random globule; *i.e.*, the number of contacts increases, the energy decreases, while the fraction of native contacts remains below 0.3. In this way, the total of $\approx 10^{16}$ random-coil conformations is effectively reduced to $\approx 10^{10}$ random semi-compact globule states. The fast collapse results from a large energy gradient (Fig. 1b) and the presence of many empty lattice points. In the second stage, which is rate limiting, the chain searches for one of the $\approx 10^3$ 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, $\tau \approx 10^{10}/(10^3 \times 1) = 10^7$, identical to the observed time scale. 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. In the third stage, the chain rapidly (within $\approx 10^5$ MC steps) attains the native conformation from any one of the transition states. The relationship of each of the three stages in the 3SRS-mechanism to other mechanisms is discussed in reference [4].

The second, rate-limiting stage is "random" in the sense that, over many trajectories, the microscopic states are occupied according to their Boltzmann probabilities and that there are many different microscopic states with comparable Boltzmann probabilities at the random-coil, random globule, and transition-state stages of folding. There is only a small difference between the folding and random non-folding sequences in the way they explore the phase space at the same absolute temperature. That is, above the glass transition temperature, the folding sequences tend to find their native states no more than an order of magnitude faster

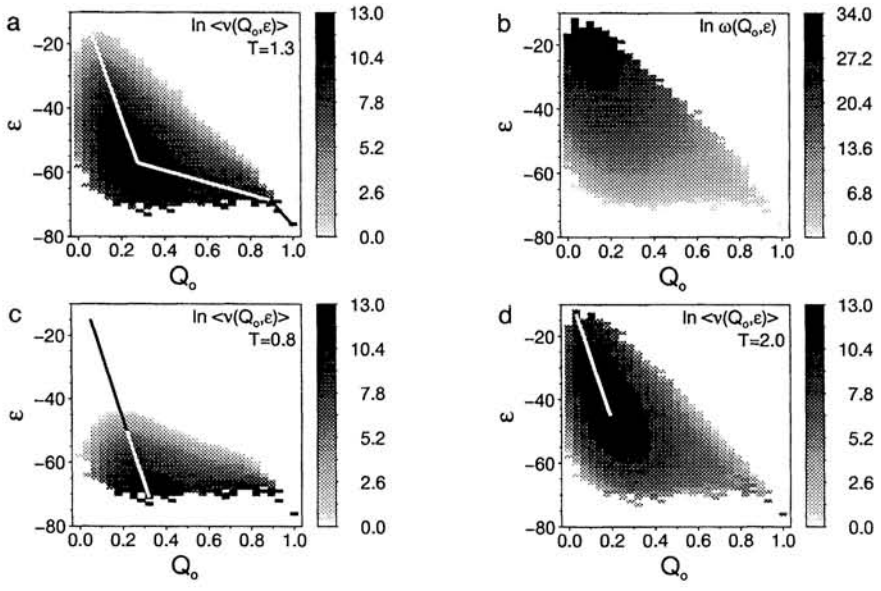

Figure 6: Density of states for a folding random sequence, as obtained from a long MC simulation. *a*, The logarithm of the average occupancy of the (Q_o, ϵ) bins, $\langle \nu(Q_o, \epsilon) \rangle$, at $T = 1.3$. Each bin spans $1/28$ Q_o units and 1 energy unit. The occupancy of a bin is the number of MC 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 $50 \cdot 10^6$ steps each [12, 54]. The broken line indicates a typical folding pathway. *b*, The logarithm of the density of states, $\omega(Q_o, \epsilon)$. The density of states is calculated from $\langle \nu(Q_o, \epsilon) \rangle$ by the use of $\frac{\langle \nu(Q_o,\epsilon) \rangle}{\langle \nu(Q_o=1,\epsilon=\epsilon_o) \rangle} = \frac{\omega(Q_o,\epsilon)}{\omega(Q_o=1,\epsilon=\epsilon_o)} \frac{\exp[-\epsilon/(k_B T)]}{\exp[-\epsilon_o/(k_B T)]}$ where ϵ_o is the energy of the native state and k_B is the Boltzmann constant set to 1; because $\omega(Q_o = 1, \epsilon = \epsilon_o) = 1$, the density of states is $\omega(Q_o, \epsilon) = \frac{\langle \nu(Q_o,\epsilon) \rangle}{\langle \nu(1,\epsilon_o) \rangle} \exp(-\frac{\epsilon_o - \epsilon}{k_B T})$. The calculation is accurate because thermodynamic equilibrium at the present Q_o, ϵ resolution is achieved, as demonstrated by a small fractional error in $\langle \nu(Q_o, \epsilon) \rangle$ (less than 10% in the populated parts of the phase space). The summation of $\omega(Q_o, \epsilon)$ over all Q_o and ϵ results in $\approx 1.1 \cdot 10^{16}$ self-avoiding chains; this is in good agreement with the extrapolation of the exact enumerations for shorter chains: $\nu^{L-1}/12 = 2.2 \cdot 10^{16}$ where L is the number of monomers, $\nu = 4.68$ is the average number of monomer states, and division by 12 corrects for symmetry [74]. Semi-compact states are defined as the states with energy less than $-45 k_B T$; summation over the appropriate region of $\omega(Q_o, \epsilon)$ yields $\approx 10^{10}$ such states. The transition states correspond to all the states with $0.8 \leq Q_o < 1$; there are $\approx 10^3$ such states. *c*, The average occupancy of the bins at a low T ($T = 0.8$) as calculated from $\omega(Q_o, \epsilon)$. 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 favored denatured states. In contrast, at $T = 1.0$, the native state occurs for $\approx 40\%$ of the time while still being kinetically accessible.

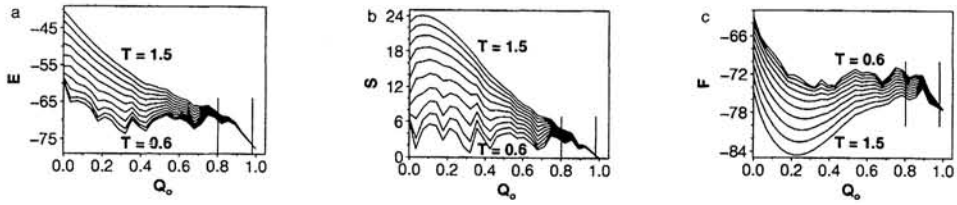

Figure 7: Reaction profiles as a function of the reaction coordinate for the folding random sequence at $0.6 \leq T \leq 1.5$. *a*, Energy, E. *b*, Entropy, S. *c*, Free energy, F. The transition state lies between the two vertical lines at $0.83 \leq Q_o \leq 0.96$. Profiles are calculated in temperature intervals of 0.1, using the partition function $Z(Q_o,T) = \sum_\epsilon \omega(Q_o,\epsilon) \exp(-\epsilon/kT)$. $E(Q_o,T) = \sum_\epsilon \omega(Q_o,\epsilon)\, p_\epsilon\, \epsilon$, $p_\epsilon = \exp(-\epsilon/k_BT)/Z(Q_o,T)$, $F(Q_o,T) = -k_BT \ln Z(Q_o,T)$, and $S(Q_o,T) = [E(Q_o,T) - F(Q_o,T)]/T$.

than the non-folding sequences [40, 52, 54]. However, in the neighbourhood of T_m, the midpoint of the folding transition, the non-folding sequences fold much slower. This is clearly different from the funnel hypothesis of protein folding which assumes that folding sequences fold because they have a single large folding funnel leading to the native conformation and that non-folding sequences do not fold because they have multiple pathways leading to several conformations [11].

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 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. While a non-folding sequence may also fold slowly to its native state above the critical temperature, it would not be stable and, therefore, could not correspond to a real protein.

Surfaces can be constructed for which resolution of the Levinthal paradox is trivial (*e.g.*, a smooth descent to an energy minimum [57], or only local interactions stabilizing the native state [66], as in the helix-coil transition). However, this does not obviate the fact that the large size of the configuration space is a necessary condition for a paradox to exist. The 27-mer model satisfies this condition with 10^{16} configurations, while short oligomers that have been extensively studied on a two-dimensional square lattice [14, 38, 39, 46] may not; for example, in reference [46], simulations involving more than 10^5 MC steps were used to fold a 13-mer on a square lattice that has only $\approx 4 \cdot 10^4$ conformers. As was pointed out previously, the shape of the configuration space, as well as the number of conformers, is important [11, 52, 66]. The landscape of the 27-mer is clearly sufficiently complex to have a Levinthal paradox; otherwise all sequences would have folded rapidly to the stable native state.

The size of the search for the native state is greatly reduced when the chain is semi-compact as it is in real proteins [67]. 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 when combined with the search of the random compact

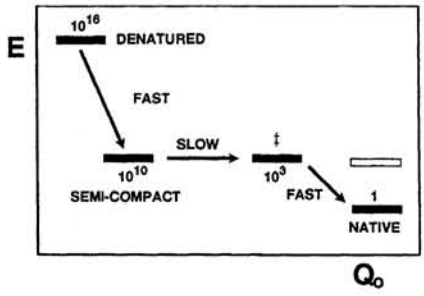

Figure 8: Three-stage random search mechanism of folding for the present model (see text). The numbers of the states were obtained from figure 6. The relative rates are obtained from 100 folding simulations (Fig. 2). The empty rectangle indicates the energy of the native state of a non-folding sequence; in the non-folding sequence, the ground state would not be populated at a temperature sufficiently high to avoid being trapped as in figure 6c. 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 mainchain conformations of which $1.7^{79} \approx 10^{18}$ are semi-compact[67]. 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, approximately 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[75], there are 52 mainchain 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 sec. This is close to the time scale observed in real proteins.

globule. Extrapolation of the folding time to real proteins suggests that the 3SRS-mechanism could be effective for small proteins (Fig. 8). However, the mechanism breaks down for long chains because the folding time is expected to increase exponentially with chain length; *i.e.*, 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 likely that proteins existing early in evolution were small enough to fold according to the 3SRS-mechanism[68]. Since 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 as high as 105°[69]. If so, the stability condition required a native state that is a very low energy minimum for which the folding problem was solved simultaneously. As evolution progressed, longer proteins evolved. These proteins had to fold on the same time scale. 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" [57] and the "principle of minimal frustration" [56]. It is also in accord with the existence of a nucleus for folding [8], or the early appearance of secondary structural elements [7, 70], neither of

which are found in the folding sequences of the present model. As a result, the collapse would not result in random semi-compact globules and the very favorable native contacts would lead more directly to a native-like molten globule state; *i.e.*, the folding pathway in figure 6a 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 [13, 34, 71]. 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 3SRS-mechanism 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.

Acknowledgements

We are grateful to Aaron Dinner, Daša Šali, Oleg Ptitsyn, Alexander Gutin, Georgios Archontis, Amedeo Caflisch, Oren Becker, and Lloyd Dimitrius for discussions concerning the protein folding problem. A.Š. was a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research. This investigation has been 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.

References

[1] T. E. Creighton (ed.), *Protein Folding*, W.H. Freeman and Company, New York, 1992.

[2] K. Merz Jr. and S. Le Grand (eds.), in *The Protein Folding Problem and Tertiary Structure Prediction*, Birkhäuser, Boston, 1994.

[3] M. Karplus and E. Shakhnovich, Protein folding: Theoretical studies of thermodynamics and dynamics, p. 127–196. In T. E. Creighton (ed.), *Protein Folding*, W.H. Freeman and Company, New York, 1992.

[4] M. Karplus and A. Šali, Theories of protein folding, *Curr. Opin. Struct. Biol. in press*

[5] C. Levinthal, How to fold graciously, in *Mossbauer Spectroscopy in Biological Systems*, Proceedings of a Meeting held at Allerton House, Monticello, IL, P. Debrunner, J. C. M. Tsibris and E. Münck (eds.), University of Illinois Press, Urbana, p. 22–24, 1969.

[6] M. Karplus and D. L. Weaver, Protein folding dynamics: The diffusion-collision model and experimental data, *Prot. Science*, 3, p. 650–668, 1994.

[7] P. Kim and R. Baldwin, Intermediates in the folding reactions of small proteins, *Ann. Rev. Biochem.*, 59, p. 631–660, 1990.

[8] D. B. Wetlaufer, Nucleation, rapid folding, and globular intrachain regions in proteins, *Proc. Natl. Acad. Sci. USA*, 70, p. 697–701, 1973.

[9] K. M. Fiebig and K. A. Dill, Protein core assembly process *J. Chem. Phys.*, 98, p. 3475–3487, 1993.

[10] E. Shakhnovich, G. Farztdinov, A. M. Gutin and M. Karplus, Protein folding bottlenecks: A lattice Monte Carlo simulation, *Phys. Rev. Lett.*, 67, p. 1665–1668, 1991.

[11] P. E. Leopold, M. Montal and J. N. Onuchic, Protein folding funnels: A kinetic approach to the sequence-structure relationship, *Proc. Natl. Acad. Sci. USA*, 89, p. 8721–8725, 1992.

[12] A. Šali, E. I. Shakhnovich and M. Karplus, How does a protein fold?, *Nature*, 369, p. 248–251, 1994.

[13] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, Specific nucleus as the transition state for protein folding: Evidence from the lattice model, *Biochem.*, 33, p. 10026–10036, 1994.

[14] C. J. Camacho and D. Thirumalai, Kinetics and thermodynamics of folding in model proteins, *Proc.Natl.Acad.Sci. USA*, 90, p. 6369–6372, 1993.

[15] E. I. Shakhnovich and A. M. Gutin, Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach, *Biophys. Chem.*, 34, p. 187–199, 1989.

[16] E. I. Shakhnovich and A. M. Gutin, Implications of thermodynamics of protein folding for evolution of primary sequences, *Nature*, 346, p. 773–775, 1990.

[17] J. D. Bryngelson and P. G. Wolynes, Intermediates and barrier crossing in a random energy model (with applications to protein folding), *J. Phys. Chem.*, 93, p. 6902–6915, 1989.

[18] A. Caflisch and M. Karplus, Molecular dynamics studies of protein and peptide folding and unfolding, *The protein folding problem and tertiary structure prediction*, K. Merz Jr. and S. Le Grand (eds.), Birkhäuser, Boston, p. 193–230, 1994.

[19] H. Frauenfelder and P. G. Wolynes, Biomolecules: Where the physics of complexity and simplicity meet, *Physics Today*, 47, p. 58–64, 1994.

[20] V. Daggett and M. Levitt, Protein folding — unfolding dynamics, *Curr. Opin. Struct. Biol.*, 4, p. 291–295, 1994.

[21] K. A. Dill, Folding proteins: finding a needle in a haystack, *Curr. Opin. Struct. Biol.*, 3, p. 99–103, 1993.

[22] H. S. Chan and K. A. Dill, The protein folding problem, *Physics Today*, 46, p. 24–32, 1993.

[23] G. D. Rose and T. P. Creamer, Protein folding: predicting predicting, *Proteins*, 19, p. 1–3, 1994.

[24] G. D. Rose and R. Wolfenden, Hydrogen bonding, hydrophobicity, packing and protein folding, *Annu. Rev. Biophys. Biomol. Struct.*, 22, p. 381–415, 1993.

[25] K. A. Dill and D. Stigter, Modeling protein stability as heteropolymer collapse, *Adv. Prot. Chem.*, in press 1995.

[26] M. Levitt, Protein folding, *Curr. Opin. Struct. Biol.*, 1, p. 224–229, 1991.

[27] R. A. Abagyan, Towards protein folding by global energy optimization, *FEBS Lett.*, 325, p. 17–22, 1993.

[28] J. Skolnick, A. Kolinski and A. Godzik, From independent modules to molten globules: Observations on the nature of protein folding intermediates, *Proc. Natl. Acad. Sci. USA*, 90, p. 2099–2100, 1993.

[29] C. L. Brooks III, M. Karplus and B. M. Pettit, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*, John Wiley & Sons, New York, 1988.

[30] M. Levitt and A. Warshel, Computer simulation of protein folding, *Nature*, 253, p. 694–698, 1975.

[31] N. Gō and H. Abe, Noninteracting local-structure model of folding and unfolding transition in globular proteins. II Application to two-dimensional lattice proteins, *Biopolymers*, 20, p. 1013–1031, 1981.

[32] J. Skolnick and A. Kolinski, Simulations of the folding of a globular protein, *Science*, 250, p. 1121–1125, 1990.

[33] K. F. Lau and K. A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins, *Macromolecules*, 22, p. 3986–3997, 1989.

[34] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations, *J. Chem. Phys.*, 101, 6052, 1994.

[35] S. Ramanathan and E. Shakhnovich, Statistical mechanics of proteins with "evolutionary selected" sequences, *Physical Review E*, 50, p. 1303–1312, 1994.

[36] A. Dinner, A. Šali, M. Karplus and E. Shakhnovich, Phase diagram of a model protein derived by exhaustive enumeration of the conformations, *J. Chem. Phys.*, 101, p. 1444–1451, 1994.

[37] J. T. Ngo, J. Marks and M. Karplus, Computational complexity, protein structure prediction, and the Levinthal paradox, *The Protein Folding Problem and Tertiary Structure Prediction*, K. Merz Jr. and S. Le Grand (eds.), Birkhäuser, Boston, p. 433–506, 1994.

[38] H. S. Chan and K. A. Dill, Transition states and folding dynamics of proteins and heteropolymers, *J. Chem. Phys.*, 100, p. 9238–9257, 1994.

[39] C. J. Camacho and D. Thirumalai, Minimum energy compact structures of random sequences of heteropolymers, *Phys. Rev. Lett.*, 71, p. 2505–2508, 1993.

[40] N. D. Socci and J. N. Onuchic, Folding kinetics of proteinlike heteropolymers, *J. Chem. Phys.*, 101, p. 1519–1528, 1994

[41] E. M. O'Toole, Effect of sequence and intermolecular interactions on the number and nature of low-energy states for simple model proteins, *J. Chem. Phys.*, 98, p. 3185–3190, 1993.

[42] A. Kolinski and J. Skolnick, Monte Carlo simulations of protein folding. I Lattice model and interaction scheme. *Proteins*, 18, p. 338–352, 1994.

[43] D. A. Hinds and M. Levitt, Exploring conformational space with a simple lattice model for protein structure, *J. Mol. Biol.*, 243, p. 668–682, 1994.

[44] D. G. Covell, Lattice model simulations of polypeptide chain folding, *J. Mol. Biol.*, 235, p. 1032–1043, 1994.

[45] M. - H. Hao and H. A. Scheraga, Monte Carlo simulation of a first-order transition for protein folding, *J. Phys. Chem.*, 98, p. 4940–4948, 1994.

[46] R. Miller, C. A. Danko, M. J. Fasolka, A. C. Balazs, H. S. Chan and K. A. Dill, Folding kinetics of proteins and copolymers, *J. Chem. Phys.*, 96, p. 768–780, 1992.

[47] O. B. Ptitsyn, Protein folding: Hypotheses and experiments, *J. Prot. Chem.*, 6, p. 273–293, 1987.

[48] M. Harding, D. Williams and D. Woolfson, *Biochemistry*, 30, p. 3120–3128, 1991.

[49] E. I. Shakhnovich and A. V. Finkelstein, Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition, *Biopolymers*, 28, p. 1667–1680, 1989.

[50] Z. - Y. Peng and P. S. Kim, A protein dissection study of a molten globule, *Biochemistry* 33, p. 2136–2141, 1994.

[51] D. A. Dolgikh, R. I. Gilmanshin, E. V. Brazhnikov, V. E. Bychkova, G. V. Semisotnov, S. Y. Venyaminov and O. B. Ptitsyn, α-lactalbumin: Compact state with fluctuating tertiary structure?, *FEBS Lett.*, 136, p. 311–315, 1981.

[52] A. Šali, E. I. Shakhnovich and M. Karplus, Kinetics of protein folding: A lattice model study of the requirements for folding to the native state, *J. Mol. Biol.*, 235, p. 1614–1636, 1994.

[53] J. Skolnick and A. Kolinski, Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics, *J. Mol. Biol.*, 221, p. 499–531, 1991.

[54] A. Šali, E. I. Shakhnovich and M. Karplus, Thermodynamics of protein folding: A lattice model study of folding to the native state, *In preparation.*

[55] J. D. Honeycutt and D. Thirumalai, The nature of folded states of globular proteins, *Biopolymers*, 32, p. 695–709, 1992.

[56] J. D. Bryngelson and P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. USA*, 84, p. 7524–7528, 1987.

[57] N. Gō and H. Abe, The consistency principle in protein structure and pathways of folding, *Adv. Biophysics*, 18, p. 149–164, 1984.

[58] R. A. Goldstein, Z. A. Luthey-Schulten and P. G. Wolynes, Optimal protein-folding codes from spin-glass theory, *Proc. Natl. Acad. Sci. USA*, 89, p. 4918–4922, 1992.

[59] M. S. Friedrichs, R. A. Goldstein and P. G. Wolynes, Generalized protein tertiary structure recognition using associative memory Hamiltonians *J. Mol. Biol.*, 222, p. 1013–1034, 1991.

[60] M. S. Friedrichs and P. G. Wolynes, Toward protein tertiary structure recognition by means of associative memory Hamiltonians, *Science*, 246, p. 371–373, 1989.

[61] M. Sasai and P. G. Wolynes, Molecular theory of associative memory Hamiltonian models of protein folding, *Phys. Rev. Lett.*, 65, p. 2740–2743, 1990.

[62] M. Sasai and P. G. Wolynes, Unified theory of collapse, folding, and glass transitions in associative-memory Hamiltonian models of proteins, *Phys. Rev. A*, 46, p. 7979–7997, 1992.

[63] E. I. Shakhnovich, Proteins with selected sequences fold into unique native conformation, *Phys. Rev. Lett.*, 72, p. 3907–3910, 1994.

[64] E. I. Shakhnovich and A. M. Gutin, Engineering of stable and fast-folding sequences of model proteins, *Proc. Natl. Acad. Sci. USA*, 90, p. 7195–7199, 1993.

[65] E. I. Shakhnovich and A. M. Gutin, A new approach to the design of stable proteins, *Prot. Eng.*, 6, p. 793–800, 1993.

[66] R. Zwanzig, A. Szabo and B. Bagchi, Levinthal's paradox, *Proc. Natl. Acad. Sci. USA*, 89, p. 20–22, 1992.

[67] K. A. Dill, Theory for folding and stability of globular proteins, *Biochemistry*, 24, p. 1501–1509, 1985.

[68] E. E. Di Iorio, W. Yu, C. Calonder, K. H. Winterhalter, G. De Sanctis, G. Falcioni, F. Ascoli, B. Giardina and M. Brunori, Protein dynamics in minimyoglobin: Is the central core of myoglobin the conformational domain?, *Proc. Natl. Acad. Sci. USA*, 90, p. 2025–2029, 1993.

[69] K. O. Stetter, Life at the upper temperature border, In *Frontiers of Life*, J. K. Trân Thanh Vân, J. C. Mounolou, J. Schneider, C. McKay (eds.), Editions Frontières, Gif-sur-Yvette, France, p. 195–212, 1992.

[70] M. Karplus and D. L. Weaver, Protein-folding dynamics, *Nature*, 260, p. 404–406, 1976.

[71] A. R. Dinner, A. Šali and M. Karplus, Folding requirements for longer model proteins, *In preparation.*

[72] S. Miyazawa and R. L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation, *Macromolecules*, 18, p. 534–552, 1985.

[73] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.*, 21, p. 1087–1092, 1953

[74] M. F. Sykes, Self-avoiding walks on the simple cubic lattice, *J. Chem. Phys.*, 39, p. 410–412, 1963

[75] R. J. Loncharich and B. R. Brooks, Temperature dependence of dynamics of hydrated

myoglobin. Comparison of force field calculations with neutron scattering data, *J. Mol. Biol.*, 215, p. 439–455, 1990