J. Mol. Biol. (1994) **235**, 1614–1636

Kinetics of Protein Folding

A Lattice Model Study of the Requirements for Folding to the Native State

Andrej Šali, Eugene Shakhnovich and Martin Karplus†

Department of Chemistry
12 Oxford St, Harvard University
Cambridge, MA 02138, U.S.A.

A three-dimensional lattice model of a protein is used to investigate the properties required for its folding to the native state. The polypeptide chain is represented as a 27 bead heteropolymer whose lowest energy (native) state can be determined by an exhaustive enumeration of all fully compact conformations. A total of 200 sequences with random interactions are generated and subjected to Monte Carlo simulations to determine which chains find the ground state in a short time; i.e. which sequences overcome the folding problem referred to as the Levinthal paradox. Comparison of the folding and non-folding sequences is used to identify the features that are required for fast folding to the global energy minimum. It is shown that successful folding does *not* require certain attributes that have been previously proposed as necessary for folding; these include a high number of short *versus* long-range contacts in the native state, a high content of the secondary structure in the native state, a strong correlation between the native contact map and the interaction parameters, and the existence of a high number of low energy states with near-native conformation. Instead, the essential difference between the folding and the non-folding sequences is the nature of the energy spectrum. The necessary and sufficient condition for a sequence to fold rapidly in the present model is that the native state is a pronounced energy minimum. As a consequence, the thermodynamic stability of the native state of a folding sequence has a sigmoidal dependence on temperature. This permits such a sequence to satisfy both the thermodynamic and the kinetic requirements for folding; i.e. the native state predominates thermodynamically at temperatures that are high enough for folding to be kinetically possible. The applicability of the present results to real proteins is discussed.

Keywords: protein folding; lattice Monte Carlo simulation; optimization

1. Introduction

A very large number of distinct conformations are available to the polypeptide chain of a protein molecule. Under physiological conditions, a protein spends most of its time in the native conformation which spans only an infinitesimal fraction of the entire configuration space. If this were not so, proteins would be of little value since they can function only if they are in the neighborhood of their native conformation. Moreover, proteins are capable of finding the unique native conformation out of the enormous number of existing conformations; this is frequently referred to as the Levinthal paradox (Levinthal, 1968, 1969). Thus, the amino acid sequence of the protein must satisfy two requirements, one thermodynamic and one kinetic. The thermodynamic requirement is that the sequence must have a unique folded conformation, which is stable and corresponds to the native structure. The kinetic requirement is that the denatured polypeptide chain can fold into this conformation under the appropriate solution conditions.

An extensive review of theoretical studies of the thermodynamics and dynamics of protein folding has been given recently (Karplus & Shakhnovich, 1992). It was pointed out that the present limitations of computing power require that simplified models be used for studies of the folding process. This is to be contrasted with the studies of the native state, which has relatively small (2 Å) and rapid (sub-nanosecond) fluctuations that have been examined in great detail with molecular dynamics

† Author to whom all correspondence should be addressed.

0022–2836/94/051614–23 $08.00/0

simulations employing potentials with a full atomic representation of the protein chain (Brooks *et al.*, 1988). Much less is known about the potential surface governing the non-native portion of conformation space involved in protein folding. It includes a wide range of structures that may differ by tens of anstroms. 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 that can serve as a basis for theoretical work on protein folding. The first 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 that can treat a molecule on the time and length scales at which protein folding occurs. This leads to an approach to protein folding that combines a simplified representation of the polypeptide chain (bead model with interactions only between spatial neighbors) with a simplified representation of the conformational space (lattice model).

Such a bead model has been used in the development of a heteropolymer theory of protein thermodynamics (Shakhnovich & Gutin, 1989, 1990*a*). The probability was estimated that a random sequence of amino acids would have a unique compact ground state and so satisfy the first requirement for a biologically relevant polypeptide chain. The probability was found to be surprisingly high with reasonable estimates of the model parameters; e.g. one percent of all random sequences in the heteropolymer model have a thermodynamically stable unique structure. This suggests that the evolutionary requirement for proteins with specific functions could have been achieved by selection from the large number of sequences with a unique ground state.

The requirement that a sequence be able to fold to the native structure in a biologically useful time may lead to more restrictive conditions. A wide range of phenomenological folding models exist. An example that has been studied in detail is the diffusion-collision model (Karplus & Weaver, 1976, 1979). Most of the models focus on the Levinthal paradox and suggest ways by which only a small fraction of the total number of conformations participate in the folding process. The existing models have been reviewed by Karplus & Shakhnovich (1992). It has been argued in the context of analogies between spin-glasses and proteins (Bryngelson & Wolynes, 1987, 1989; Shakhnovich & Gutin, 1989, 1990*a*) that the energy surface of a protein is "rugged". This means that there are many energy barriers that have to be crossed during the folding process. On such a surface, folding requires that there exists a temperature high enough for the process to occur (i.e. the protein is not frozen in one of the minima) yet low enough so that the ground state is stable.

One way of shortening the folding time on a complex energy surface with many local minima is to introduce a hierarchy into the process; e.g. secondary structure forms first, followed by coalescence of the secondary structural elements to yield the tertiary fold of the native state. The diffusion-collision model (Karplus & Weaver, 1976, 1979) and the closely related framework model (Kim & Baldwin, 1982, 1990) embody this idea. Additionally, it has been suggested that the entire surface may be significantly biased toward the native state so as to channel the folding process. Such biases have been introduced by the choice of the energy function that makes only native contacts stabilizing (Gō & Abe, 1981) or by introducing a weighting function that favors the native backbone dihedral angles (Bryngelson & Wolynes, 1987; Skolnick & Kolinski, 1990). It has been shown in lattice simulations that either of these biases can be sufficient for rapid folding. Both of these assumed biases have been related to the "principle of minimal frustration" (Bryngelson & Wolynes, 1987, 1989), which postulates that the energy minimum of a small part of a protein in isolation is not very different from the structure of that part in the native state of the entire protein.

Although the above models may be sufficient to achieve rapid folding, little has been done to show that they are necessary or that they occur in real proteins. Moreover, local dihedral angle biases, which reduce protein folding to an analog of the helix-coil transition (Zwanzig *et al.*, 1992), do not lead to a cooperative (all-or-none) folding transition, one of the essential experimental aspects of protein thermodynamics (Karplus & Shakhnovich, 1992). Instead, longer-range interactions, which are an essential aspect of the complexity of folding (M. Karplus & E. Shakhnovich, unpublished results), are required for cooperativity. Thus, two major unresolved conceptual problems in understanding the mechanism of protein folding are the determination of the necessary conditions for overcoming the Levinthal paradox and the demonstration of which mechanism is employed in the folding of polypeptide chains. In the present study, we attempt to examine these problems by the use of a simplified lattice model.

Since we are unable to use an analytical approach, as it has been done for the thermodynamic aspect of folding (Shakhnovich & Gutin, 1989, 1990*a*), we make use of an unbiased lattice model that possesses a very large number of conformations that cannot all be scanned in a folding simulation, and has a well defined native state. Although lattice models have been used by others (see the review by Karplus & Shakhnovich, 1992), the present approach differs in two essential aspects. First, we choose a system where the lowest energy state is known, unique, and satisfies the thermo-

dynamic condition on the sequence that can correspond to a functioning protein. This is achieved by introducing an overall (hydrophobic) compactness condition on the polypeptide chain so that the low energy portion of the conformation space can be searched exhaustively; i.e. all compact self-avoiding conformations can be enumerated and the lowest energy state determined from this enumeration. Second, rather than introducing preconceived biases into the energy function, we use an unbiased energy function that consists of contact interactions between beads chosen from a Gaussian distribution. We generate a series of random sequences and determine which of them can fold to the stable native state in a reasonable time using a Monte Carlo search procedure. In this way, we obtain an unbiased sample of folding and non-folding sequences. The differences between the two sets allow us to obtain insights into the kinetic condition for a biologically relevant polypeptide sequence.

The specific protein model is a 27 bead heteropolymer chain on a cubic lattice. This model has already been used in a preliminary study of the folding kinetics (Shakhnovich *et al.*, 1991). Despite its simplicity, the model contains certain essential features of protein folding. The total number of conformations of a 27-mer on a cubic lattice, including the non-compact conformations, is approximately 5^{26}, a number much larger than can be scanned by any simulation. Thus, the Levinthal paradox is present in this system. Successful folding can occur only if the energy surface is such that the molecule avoids most of the conformations in the folding process. Moreover, comparison of thermodynamic results obtained by numerical simulations of a 27-mer with those from the analytical theory for infinite heteropolymers (Shakhnovich & Gutin, 1989, 1990*a*) has shown that a 27-mer chain is a good model for the thermodynamics of long chains. This suggests that the kinetic results obtained with the 27-mer model are likely to be applicable to longer chains.

The procedure followed when using this model to study the kinetics of folding is to generate a large number of random sequences. Starting with a coil configuration, these sequences are subjected to a Monte Carlo simulation under conditions where the native state is stable to determine whether or not they fold in a reasonable number of steps. This provides a data base of sequences that do and do not satisfy the thermodynamic and kinetic criteria for a functional sequence. Definition of successful folding is the most complicated part, but it is essential. Since the unique ground state is known, it is possible to make a rigorous determination of whether a given simulation reaches the ground state; a successful folding simulation is one that finishes at the global energy minimum. Correspondingly, "folding" sequences are those that can find this global minimum a number of times, independent of the initial conformation; the non-folding sequences are those that cannot find it in a reasonable time. Thus, the present simulations avoid the problems that occur when the global energy minimum is not known: it has often been observed that folding is dependent on initial conditions and that a simulation may terminate in a metastable trap (Honeycutt & Thirumalai, 1992).

We analyze the results to find the features of the folding sequences that distinguish them from the non-folding sequences. We use the model to explore the relation between folding and (1) characteristics of the energy spectrum, (2) structural features of the native conformation, and (3) the relationship between the energy levels and corresponding conformations. We also discuss the relevance of our results for the folding of real proteins, for methods that predict protein three-dimensional (3D†) structure, and for general stochastic optimization methods.

Section 2 describes the methods used in the calculations. The results are presented in Section 3 and discussed in Section 4.

2. Methods

(a) *Lattice model*

The numerical simulations in this paper use short chains of 27 monomers with discrete positions (Shakhnovich *et al.*, 1991: Fig. 1(a)). This choice is made to satisfy the full enumeration condition. The bonds between monomers all have unit length. Allowed monomer positions include only cubic lattice sites. Contacts can be formed only between 2 monomers that are not successive in sequence and are at unit distance from each other. There are at most 5 such contacts for the 2 terminal monomers and 4 for the other monomers. Thus, a fully compact self-avoiding chain corresponds to a $3\times3\times3$ cube with 28 contacts. There are 103,346 such structures unrelated by symmetry (Shakhnovich & Gutin, 1990*b*). When multiple conformations related by symmetry are not excluded, there are 4,960,608 compact self-avoiding structures (Chan & Dill, 1990). Additionally, all non-compact self-avoiding conformations can occur during folding process; there are approximately $5^{26} \approx 10^{18}$ such conformations.

The energy of the polymer chain is assumed to depend only on nearest neighbor contacts and to be independent of other aspects of the chain conformation (e.g. there are no pseudo-dihedral angle biases). The energy function is taken to have the simple form:

$$E = \sum_{i<j} B_{ij}\Delta(r_i - r_j), \tag{1}$$

where B_{ij} is the interaction energy between monomers i and j located at positions r_i and r_j, respectively. $\Delta(r_i - r_j)$ is 1 if monomers i and j are in contact and is 0 otherwise. To permit the study of a model without preconceived biases, the interaction parameters B_{ij} are obtained from a Gaussian distribution with a mean B_o and standard deviation σ_B; that is,

$$P(B_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_B} e^{-\frac{1}{2}\left(\frac{B_{ij}-B_o}{\sigma_B}\right)^2}. \tag{2}$$

† Abbreviations used: 3D, three-dimensional; 2D, two-dimensional; REM, random energy model.

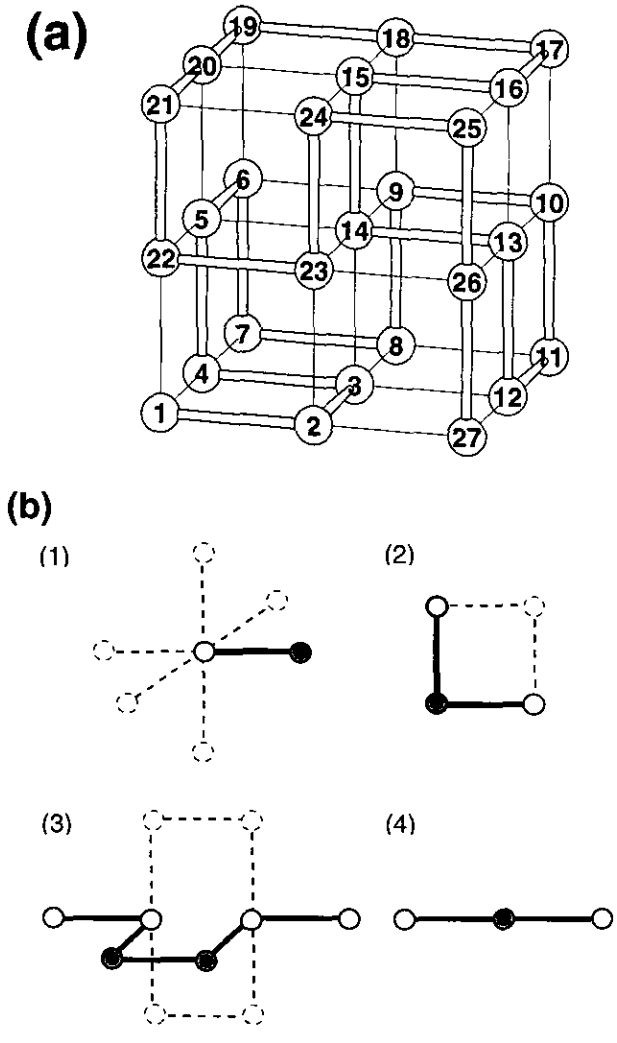

Figure 1. The lattice model of protein folding. (a) The cubic lattice and an example of a compact self-avoiding walk of a chain (thick line) of 27 monomers (numbered in order); the native structure of sequence 43 is shown. (b) The 3 types of possible Monte Carlo moves (1 to 3; see Methods). Situation 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 broken lines. A move is possible if all new positions are unoccupied. The monomers that are moved are shown in dark stippling.

This model for B_{ij} corresponds to a heteropolymer with a random sequence of monomers of many different types. The parameter σ_B measures the degree of heterogeneity of a chain; a homopolymer with all monomers interacting with the same energy has $\sigma_B = 0$. As B_0 becomes more negative, collapse from a random coil to a globule is favored. The parameter B_0, which introduces the overall attractive term, emulates the hydrophobic effect observed in globular proteins. When B_0 is sufficiently negative, the global energy optimum generally corresponds to one of the compact self-avoiding conformations. In this limit, only 103,346 conformations have to be checked to find the global energy optimum. Equations (1) and (2) provide a complete description of the heteropolymer being considered here. This choice is of particular interest because it is the starting point of the analytical theory and derivation of the phase diagram of a random heteropolymer by Shakhnovich & Gutin (1989).

The Monte Carlo simulation of protein folding starts with a random self-avoiding conformation. Such a conformation corresponds to an extended denatured chain because of the much higher probability of a random process generating such a state; the number of contacts is in the range of 5 to 10. This conformation is then iteratively updated by a large number of small discrete changes. Each Monte Carlo step consists of the following: a move is selected at random until a move is found that conserves the unit bond lengths and does not result in more than 1 monomer per lattice site (a self-avoiding walk). A comparison of the present results with those obtained when multiple occupancy is allowed (Shakhnovich *et al.*, 1991) is given in Appendix II. Once an allowed move is found, the corresponding energy change in the system, ΔE, is evaluated (eqn (1)) and the Metropolis criterion is used to accept or reject the proposed move (Metropolis *et al.*, 1953). The Metropolis criterion involves a comparison of $\exp(-\Delta E/k_B T)$ with a random number uniformly distributed between 0 and 1. The move is rejected if the random number is larger than the exponential. All moves that decrease the energy of the system are accepted. T is the temperature of folding. Each application of the Metropolis criterion is counted as one Monte Carlo step (Metropolis *et al.*, 1953).

Moves (Fig. 1(b)) are proposed in the following way. First, a decision is made whether positions of 1 or 2 monomers will be changed in the current Monte Carlo step: a move of a single monomer is selected with a probability of 0·2. Moves involving more than 1 particle are included because it has been shown that the kinetics on a lattice with only single particle moves and with excluded volume constraints is unrealistic (Hilhorst & Deutch, 1975). If a single monomer move is selected, a monomer index is chosen randomly from 1 to N. For terminal monomers, 1 of the 5 possible moves is then picked randomly (move type 1). For internal monomers, there is only 1 new position if monomers $i-1$, i, $i+1$ form a right angle (move type 2). If they are on a line there is no new position and the selection of a move starts from the beginning. When a move of 2 monomers is selected, a monomer index is chosen randomly from 1 to $N-3$. If monomers i to $i+3$ form a crankshaft (i.e. the distance between i and $i+3$ is 1) 1 of the 2 possible 90° crankshaft rotations is picked randomly; otherwise, a move is unsuccessful and the selection of a move starts from the beginning.

For single and double monomer moves, about 8% and 0·5%, respectively, of the moves tested by the Monte Carlo criterion are accepted for the trial sequence in the initial phases of folding (sequence 2, $T = 1$, $B_0 = -2$). The number is much smaller for collapsed structures. The ratio of single *versus* double monomer moves tested by the Monte Carlo criterion is approximately 5; this is so because the crankshafts are rare and because a crankshaft move is likely to result in double occupancy.

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 1, backbone conformation is:

$$X(T) = 1 - \sum_i^M p_i^2, \tag{3}$$

where:

$$p_i = \frac{\exp(-E_i/k_B T)}{Z}, \qquad Z = \sum_i^M \exp(-E_i/k_B T). \tag{4}$$

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, and Z is the chain partition function. If there is a unique ground state with a large Boltzmann weight ($p_0 \sim 1$, $p_{i>0} \sim 0$), $X \sim 0$. If many states have comparable occupation ($p_i^2 \ll 1$), $X \sim 1$.

In the analytic random heteropolymer model corresponding to eqns (1) and (2), it is found that the average of X over all possible random sequences (Shakhnovich & Gutin, 1989) is:

$$\langle X \rangle = \begin{cases} T/T_c & \text{if } T \leq T_c \\ 1 & \text{if } T > T_c \end{cases} \quad (5)$$

where T_c is the critical temperature given for the ensemble of all random sequences by:

$$T_c = \frac{\sigma_B \sqrt{\rho}}{2k_B \sqrt{\ln \gamma}}. \quad (6)$$

γ is the number of conformations per monomer and ρ is the average number of contacts per monomer (each contact is counted only once, as in eqn (1)). Thus, at sufficiently low temperatures, $\langle X \rangle$ approaches zero; this means that the ground state thermodynamically dominates the distribution of states. The fraction of random sequences that have the native state with Boltzmann probability p_0 of at least $1-\varepsilon$ is (Shakhnovich & Gutin, 1990*a*):

$$P(p_o \geq 1-\varepsilon) = \frac{\sin(\pi\langle X \rangle)}{\pi\langle X \rangle} \varepsilon^{\langle X \rangle}. \quad (7)$$

The number of different conformations of a heteropolymer chain is γ^{N-1}. The number of thermodynamically relevant conformations of a given sequence, N_{TD}, can be calculated in the random heteropolymer model (Shakhnovich & Gutin, 1989). For the limiting values of X,

$$N_{TD} = \exp\left(\frac{S}{k_B}\right) \approx \begin{cases} \exp \dfrac{1}{1-X} & \text{if } X \approx 1 \\ 1+X & \text{if } X \ll 1 \end{cases} \quad (8)$$

where S is the configurational entropy of a chain. Correspondingly, $X(T)$ reflects the degree of conformational heterogeneity of a single sequence at temperature T.

A convenient measure of the similarity between two conformations i and j is the fraction of common contacts between them, Q_{ij}. We can calculate the probability distribution of Q for an ensemble of conformations of a given sequence at temperature T:

$$P_T(Q) = \sum_{i,j}^{M} \delta(Q - Q_{ij}) p_i p_j, \quad (9)$$

where $\delta(Q-Q_{ij}) = 1$ if $Q = Q_{ij}$ and is 0 otherwise. Note that $P_T(1) = 1 - X(T)$. For random heteropolymers, $\langle P(Q) \rangle$ has a bimodal shape. This is found analytically in the limit of infinite length chains and also for the cubic lattice model of a 27-mer (Shakhnovich & Gutin, 1989, 1990*a*). The first peak at $Q = 1$ corresponds primarily to the overlap of the native state with itself; it also has contributions from the self-overlap of other low energy structures. The second peak at $Q \sim 0$ (0·3 in the 27-mer) is dominated by the overlap between the native state and other low energy states. This peak is increased by the overlap between the relatively stable non-native states and other non-native states unrelated to them. The fact that Q is as small as 0·3 implies that there is little relation between the geometry of the native state and other low energy states. An energy surface with this property is referred to as "rugged" or a "rugged landscape" (Frauenfelder *et al.*, 1991; Karplus & Shakhnovich, 1992).

(b) *Definitions and choice of variables*

A particular sequence is defined by the matrix **B** of interaction parameters B_{ij}. The native conformation is the compact self-avoiding chain with the lowest energy. A sequence folds in a given Monte Carlo simulation if it finds the native conformation within a reasonably small number of Monte Carlo steps (see below); the conditions of the folding simulations are such that it would not necessarily remain in the native conformation if the simulation was continued (Shakhnovich *et al.*, 1991). *Foldicity* of a given sequence is defined as the fraction of all Monte Carlo runs that started with a random conformation and finished in 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 under conditions where the native structure is thermodynamically stable.

To perform the Monte Carlo simulations, several parameters have to be chosen. These parameters are the maximal number of steps in a simulation, the temperature T, the collapse parameter B_o, and the heterogeneity parameter σ_B. In the work by Shakhnovich *et al.* (1991), the values of these parameters were chosen by trial-and-error. Their values were used in this paper as a starting point for refinement. The maximal number of Monte Carlo steps was set to 50×10^6 because folding is usually achieved within 10×10^6 steps. The simulation was stopped when the native state was reached for the first time. This corresponds to measuring the first passage time and is sufficient for our aim of studying the kinetics of folding. An exploration of equilibrium properties would require simulations to extend beyond the first passage. Since only the relative values of the parameters B_o, σ_B and T are important, the heterogeneity parameter σ_B was fixed at 1. Folding of a trial sequence (sequence 2 in Results, section (a)) was explored to find folding parameters for the simulations of all sequences (Fig. 2). Foldicity of sequence 2 is optimal in a relatively wide range of B_o values (Fig. 2(a)). However, it is advantageous to use the most negative value of B_o ($B_o = -2$) that still gives significant folding to increase the probability that the global energy optimum corresponds to a compact self-avoiding conformation (Results, section (c)). When $B_o = -2$ and $\sigma_B = 1$, almost all B_{ij} are less than 0 (eqn (2)).

Foldicity of sequence 2 is maximal in a broad range of temperatures where $X^{csa}(T)$ is larger than 0·5 and smaller than 0·9 (Fig. 2(b); insert); the superscript csa indicates that only the compact self-avoiding chains were used in the calculation of X (M in the summation of eqn (3) is 103,346). The temperature used for the subsequent folding simulations was obtained from two considerations: (1) thermodynamic stability of the native state must be ensured and (2) foldicity should be as high as possible. The first requirement is best satisfied by low temperatures and imposes an upper limit on the folding temperature. For the 200 sequences studied, this limit corresponds to a temperature at which $X^{csa} = 0{\cdot}8$ because the native conformation still has a relatively high average Boltzmann weight (0·4) at this temperature (Fig. 3(a)); all other conformations have probabilities $\leq 0{\cdot}2$. Note that non-compact conformations are not taken into account in these calculations and that the true thermodynamic stability is less than 0·4 (see also the Discussion). The

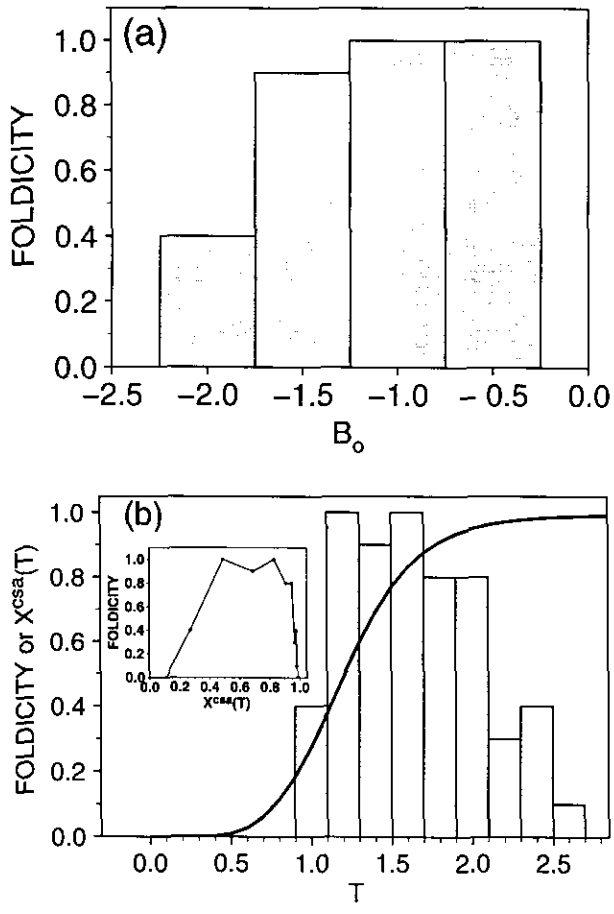

Figure 2. Optimal folding conditions for the trial folding sequence 2. Foldicity is determined from 10 independent Monte Carlo folding simulations. (a) The dependence of foldicity on the collapse parameter B_o at $T = 1{\cdot}0$. (b) The histogram shows the dependence of foldicity on temperature T at $B_o = -2{\cdot}0$. The curve is $X^{csa}(T)$ on the same scale. The inset shows the dependence of foldicity on $X^{csa}(T)$.

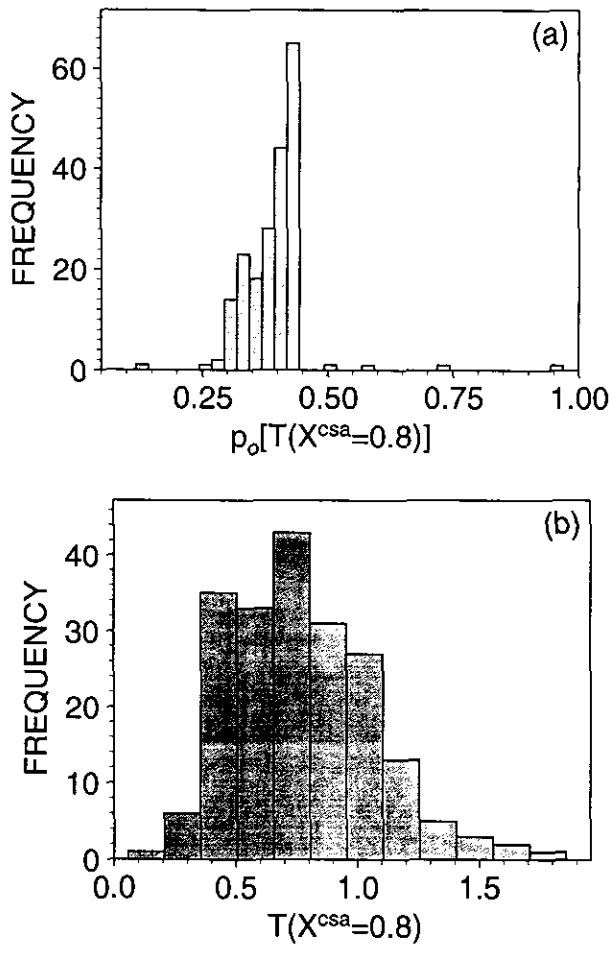

Figure 3. Determination of folding temperature. (a) Distribution of Boltzmann probabilities of the native structure for the 200 random sequences at $T(X^{csa} = 0{\cdot}8)$. The Boltzmann probabilities were calculated using an ensemble of compact self-avoiding chains only. (b) Distribution of $T(X^{csa} = 0{\cdot}8)$ for the 200 random sequences in the database.

second requirement is generally expected to be better satisfied by temperatures higher than $T(X^{csa} = 0{\cdot}8)$, especially for the sequences that have low $T(X^{csa} = 0{\cdot}8)$; the trial sequence 2 has one of the highest $T(X^{csa} = 0{\cdot}8)$ values among the 200 sequences studied. A higher temperature increases the probability of a system being able to overcome the energy barriers on its way to the native state. Thus, as a compromise between the 2 requirements, the folding temperature for all sequences was set to the highest value that reasonably fulfills the stability condition; i.e. the absolute temperature was selected at which $X^{csa} = 0{\cdot}8$. Defining a relative temperature τ:

$$\tau = \frac{T}{T_x}, \quad T_x = T(X^{csa} = 0{\cdot}8), \tag{10}$$

we use $\tau = 1$. The temperature used for folding is therefore different for each sequence. For the energy functions describing the heteropolymer (eqns (1), (2)), T_x is distributed approximately normally with a mean of 0·7 and standard deviation of 0·3 (Fig. 3(b)).

Use of constant τ with $\tau = 1$ rather than constant T, is justified as follows. The generation of the B_{ij} with eqn (2) yields a set of protein sequences with different T_x and different $X^{csa}(T)$. For sequence 2, $X^{csa}(T)$ shows the folding transition region (Fig. 2(b)). A real protein would have a much steeper transition than the 27-mer. By using $\tau = 1$, we are studying the folding kinetics in the same portion of the transition region for all sequences. If a high constant T were used, some sequences would not satisfy the thermodynamic stability criterion; i.e. they would be denatured. If a low constant T were used, many sequences would cease to fold because they would not have enough energy to overcome the energy barriers. Use of an appropriate constant τ ensures that the native state is stable and that a sequence still has a possibility of folding.

The dependencies of folding on T and B_o shown in Fig. 2 for the trial sequence were found to be typical for other folding sequences (results not shown).

When several groups of sequences are compared (e.g. folding *versus* non-folding sequences), the standard error of the mean:

$$\sigma_e(\langle x \rangle) = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(x - \langle x \rangle)^2}{n}}, \tag{11}$$

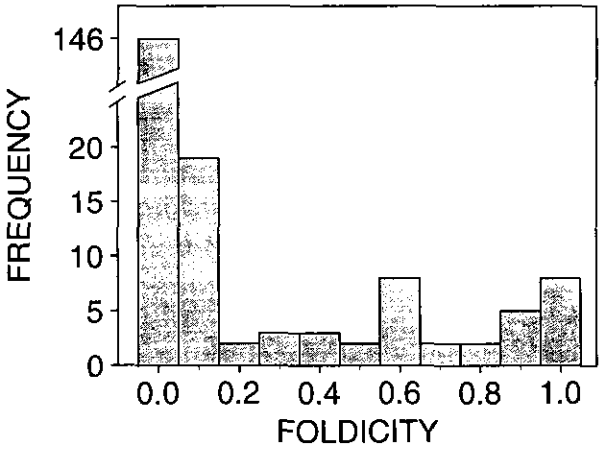

Figure 4. Distribution of foldicities for the 200 random sequences. See Methods for folding conditions.

is used to judge the significance of the differences between the groups where x is the quantity compared and n is the number of the sequences in the group.

We can compare the energy function for the current lattice model with that for the solvated proteins. We do this by comparing the average residue–residue contact energies. For solvated proteins, we use the residue–residue contact energies e_{ij} listed in Table V of Miyazawa & Jernigan (1985). Their expression for the total energy difference between the solvated extended and solvated native structure of a protein is the same as that for the lattice model (eqn (1)), except that e_{ij} values are used instead of B_{ij} values. Thus, the proper averages of e_{ij} and B_{ij} should be directly comparable. Residue–residue contacts in a protein are assumed to occur when the centers of the 2 side-chains are closer than 6·5 Å. In a typical globular protein of 100 to 200 residues, there are between 1·7 and 2·0 contacts per residue (Miyazawa & Jernigan, 1985). This can be contrasted with 1·04 (28/27) contacts per monomer for a lattice model. For proteins, the average contact energy is $-2{\cdot}64\ k_BT$ and the standard deviation for the contact energies is $1{\cdot}37\ k_BT$. In this averaging, e_{ij} values are weighted according to the frequencies of the corresponding residue types in the protein database (column 2 in Table III, Miyazawa & Jernigan, 1985). For the 30 folding sequences in the lattice model, the average contact energy is $\langle B_{ij}\rangle = -2{\cdot}02(\pm 0{\cdot}01)$, the standard deviation of B_{ij} values is $\sigma_B = 1{\cdot}00$, and the average folding temperature is $\langle T_x\rangle = 1{\cdot}24(\pm 0{\cdot}035)$. This corresponds to an average contact energy of $-1{\cdot}6\ k_BT$ and to the standard deviation of $0{\cdot}81\ k_BT$. Thus, the values for real proteins in the Miyazawa & Jernigan model are only a factor of 1·7 larger than the values for the lattice model. Moreover, there is also agreement between the lattice model and real proteins. Temperature denaturation experiments (Privalov, 1989; Privalov & Khechinashvili, 1974) suggest that the average residue–residue contact energy of myoglobin at 303 K is $1{\cdot}24\ k_BT$. Similar values are obtained for other proteins (Privalov & Khechinashvili, 1974). We note that this approximate agreement of the lattice model with the *Miyazawa & Jernigan model and real proteins was* obtained solely by considering the B_o, σ_B and T that result in optimal folding in the lattice model, not by optimizing the match between the 2 sets of parameters.

It should be noted that Miyazawa & Jernigan (1985) did not use the correct temperature with the Boltzmann-like statistics employed to calculate the energies from the observed contact frequencies. They used room temperature, but it has recently been suggested that they should have used the critical temperature T_c (Gutin *et al.*, 1992). Gutin *et al.* (1992) estimated that the critical temperature T_c is a factor 1·5 higher than room temperature. This would result in an increase of the corrected contact energies of Miyazawa & Jernigan for the same factor. However, the results in this paper indicate that rapid folding into the stable native structure occurs slightly above the critical temperature T_c (see Discussion). Because the folding temperature can be equated with room temperature, the critical temperature could also be less than room temperature. Whatever the precise value of the critical temperature, the difference between the room temperature and the critical temperature is small and does not destroy the approximate agreement between the contact energies in the Miyazawa & Jernigan model and in the lattice model.

3. Results

(a) *Database for analysis*

A total of 200 random **B** matrices were generated by using random numbers and the Gaussian distribution in equation (2). Each of these sequences was subjected to ten independent Monte Carlo simulations starting with a random coil state to determine which of the sequences fold to the ground state. The conditions described in Methods, section (b), were used. The distribution of foldicities, defined in the same section, is shown in Figure 4. There are 30 strongly folding sequences that folded four or more times, 24 weakly folding sequences that folded between one and three times, and 146 sequences that did not fold at all. The division of the folding sequences into strongly and weakly folding subgroups is useful because the trends described below are more pronounced if weakly folding sequences are omitted from the analysis.

(b) *Test of metastability of folded states*

It is conceivable for a polymer chain to fold repetitively into a unique local energy minimum that is different from a global minimum. This could be caused by kinetic "funnels" on the configuration surface (Leopold *et al.*, 1992). In fact, the question has often been raised as to whether the native structure of a protein corresponds to a free energy minimum (Honeycutt & Thirumalai, 1992; Levinthal, 1968; Wetlaufer & Ristow, 1973), though a range of experiments, at least for small proteins, suggest but do not prove that it is (Anfinsen, 1973; Karplus & Shakhnovich, 1992). In principle, protein structures could correspond to metastable states because any unique, stable and rapidly accessible structure would be sufficient for a real protein to fulfill its biological role.

Because we have a database of random sequences that do not find the ground state, it is possible to determine whether any of them do, in fact, end up in a single metastable state. To test the frequency of repetitive folding into a unique local minimum, the

conformations with the lowest energy from the ten independent folding simulations were compared for each of the 146 non-folding sequences. They all correspond to relatively low energy minima; e.g. their energies were about ten units above the ground state energies. None of these sequences folded into the same local energy minimum more than once. The fraction Q of common contacts between these structures is in the range of 0·2 to 0·4; this corresponds to the similarity expected from a set of randomly chosen structures. Therefore, the present model does not support the metastable native state model. This is in accord with the results of Honeycutt & Thirumalai (1992) who have done off-lattice simulations. They also found that the system ended up in different metastable states.

(c) *Comparison of lower parts of energy spectrum of the ensemble with compact self-avoiding chains*

The ensemble of all self-avoiding chains, which includes many more non-compact than compact structures (Chan & Dill, 1990), contains too many conformations to find the global energy minimum by evaluating each one of them. The ensemble is also too large to calculate thermodynamic functions such as X by using the summation over all structures. Instead, only the compact self-avoiding chains are considered in this paper. To determine the validity of this simplification, a Monte Carlo simulation of 50×10^6 steps was performed for all chains to find the global energy minimum. The simulation started from the native structure (i.e. the self-avoiding structure with the lowest energy) at a high temperature of $T = 1{\cdot}2$. Only 11 out of the 200 random sequences were found to have an energy minimum more negative than that of the native structure. The same Monte Carlo simulation was used to explore the lower discrete part of the energy spectrum of the 200 random sequences. The sample energy spectra of ten random sequences (Fig. 17(a)) show that if the discrete part of the spectrum is sparse for compact self-avoiding chains, it is likely to be sparse for all chains. Therefore, there is a strong correlation between, for example, X and X^{csa} (Fig. 17(b)). These two results imply that the features that depend only on the lower discrete part of the spectrum can be characterized by use of the compact self-avoiding chains alone, neglecting the non-compact conformations. Full enumeration of shorter chains in 2D confirms this finding (A. Dinner, A. Šali, E. Shakhnovich & M. Karplus, unpublished results). It is these features that are dominant in discriminating the folding from non-folding sequences (see below).

(d) *Relation between foldicity and the energy spectrum*

Since the heteropolymers are characterized by the contact energy distributions (eqns (1) and (2)) independent of the nature of conformations, it is of interest to determine whether there is a relation between foldicity and some features of the distribution of the energy levels of compact self-avoiding chains. The energy spectra for four folding and six non-folding sequences are shown in Figure 5. It is apparent that folding is associated with a large gap between the lowest and second lowest energy levels of compact self-avoiding chains, ΔE_{10}.

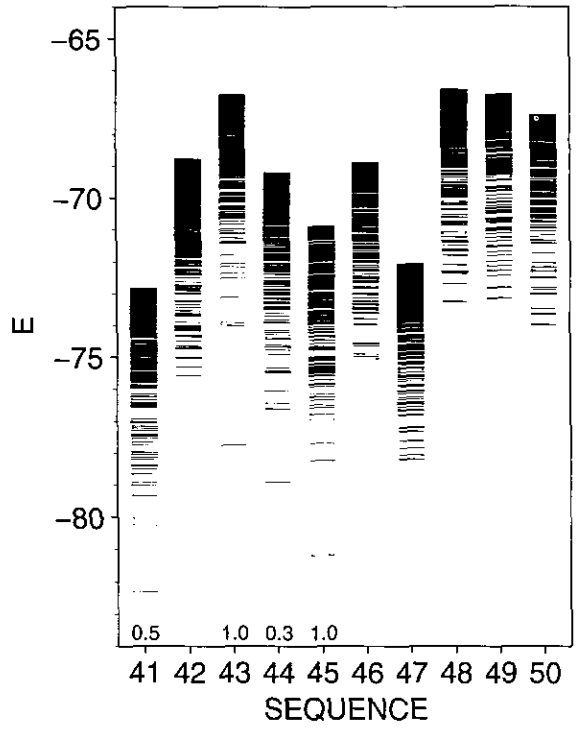

Figure 5. Energy spectra for 10 sample folding and non-folding sequences. The energies of the lowest 400 compact self-avoiding conformations are drawn. The numbers below the spectra show the foldicities of the corresponding sequences; if no number is given, the sequence does not fold.

Average $X^{csa}(T)$ for strongly folding and non-folding sequences are shown in Figure 6. The folding sequences are characterized by a sigmoidal shape of $X^{csa}(T)$ while the non-folding sequences conform to $X^{csa}(T)$ found for random chains by Shakhnovich & Gutin (1989) (Fig. 17(b)). In fact, the former corresponds to a cooperative folding transition while the latter does not. The difference in the form of the curves can be related directly to the nature of the lower part of the energy spectrum. A sparse spectrum with a large ΔE_{10} leads to a cooperative curve, while a quasi-continuous low energy spectrum results in the non-cooperative behavior. It is clear that if ΔE_{10} is larger, a higher T will be required to excite the system into energy levels above the ground state at equilibrium. This is reflected in the slower increase in $X^{csa}(T)$. Associated with this is the increase in T_x, the value at which $X^{csa}(T) = 0{\cdot}8$ (eqn (10)); $\langle T_x \rangle$ for the folding sequences is 1·24 while $\langle T_x \rangle$ for the non-folding sequences is 0·63.

The dependence of foldicity on T_x and ΔE_{10} is

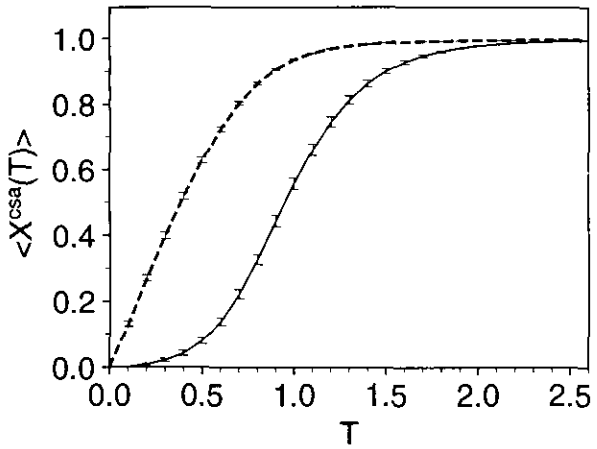

Figure 6. $\langle X^{csa}(T)\rangle$ for strongly folding (continuous line) and non-folding (broken line) sequences at $\tau = 1$. The error bars show the standard error of the mean.

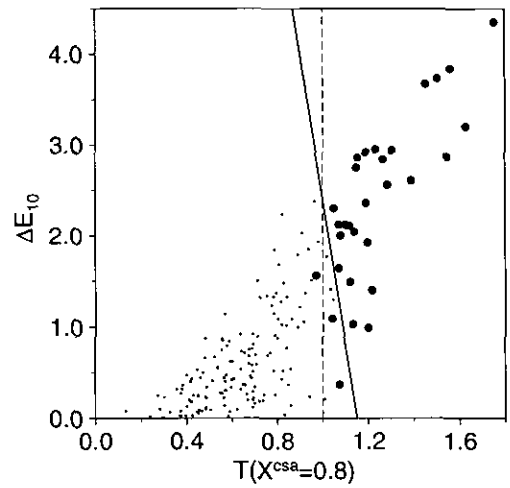

Figure 7. Discrimination between strongly folding and non-folding sequences on the basis of the energy gap, ΔE_{10}, and T_x. Strongly folding sequences, filled circles; non-folding sequences, dots. The weakly folding sequences (not shown) are concentrated around the boundary between the strongly folding and non-folding sequences. They were omitted from the plot in order to obtain better contrast between folding and non-folding sequences. The continuous line separates the 2 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 broken line indicates the critical temperature T_c for the ensemble of compact self-avoiding chains estimated from Fig. 17(b) using eqn (5).

shown in Figure 7. It confirms that ΔE_{10} and T_x are strongly correlated. Both of these parameters of the energy spectrum determine whether or not a given sequence is a folding sequence. A line can be drawn that separates the $(T_x, \Delta E_{10})$ space into two regions corresponding to the folding and non-folding sequences. When the line is drawn to minimize the number of the non-folding and folding sequences in the folding and non-folding part of the plot, respectively, only three false non-folding sequences are predicted out of the total sample of 200 random sequences (Fig. 7); the line is given by the equation $\Delta E_{10} = 18 - 16 \times T_x$. That the dividing line is almost vertical, demonstrates the importance of T_x. An optimally placed horizontal line would also separate most folding from the non-folding sequences, indicating that ΔE_{10} by itself is also significant. Since the latter (ΔE_{10}) is a simpler criterion than the former (T_x), we consider both in what follows. The high prediction success for the optimal line shows that it is possible to use only the two simple thermodynamic parameters to identify almost all strongly folding sequences without performing kinetic Monte Carlo simulations.

(e) *Associations between foldicity and the conformation of the native state*

The order of a contact between two monomers is defined as the absolute difference between their residue indices: $\Delta i = |i-j|$ with $\Delta i \geq 3$. Due to the geometric properties of the cubic lattice, only contacts of an odd order are possible. A total of 156 ($2^{\frac{12+1}{2}} \times 12$) different possible contacts exist for a 27-mer; each compact self-avoiding conformation has 28 of these.

It has been suggested that stabilizing contacts between residues close in sequence govern nucleation events which would greatly accelerate the folding process (Wetlaufer, 1973). If this were true the frequency of local contacts in the native structure of a folding sequence would be expected to be higher than that in the non-folding sequences. The distribution of spatial contacts as a function of their order for native structures of the folding and non-folding sequences, and for all 103,346 compact self-avoiding chains is shown in Figure 8. While local contacts are more likely than global contacts, as

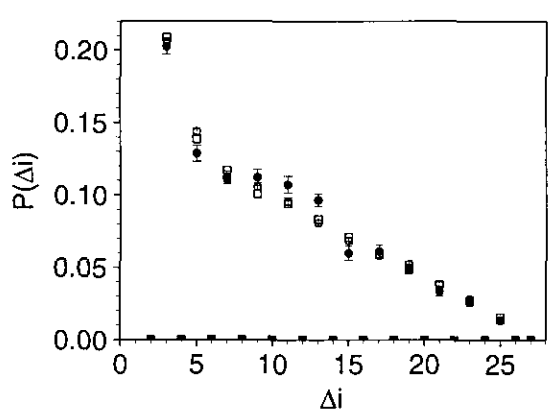

Figure 8. Distribution of contacts in native structures of strongly folding and non-folding sequences, and in all compact self-avoiding chains. Δi is the order of a contact, defined as the absolute difference of indices of monomers in contact. Strongly folding sequences, filled circles; non-folding sequences, open circles; compact self-avoiding chains, squares. The error bars show the standard error of the mean. The points at the bottom are an artefact of the cubic lattice resulting in no contacts of an even order.

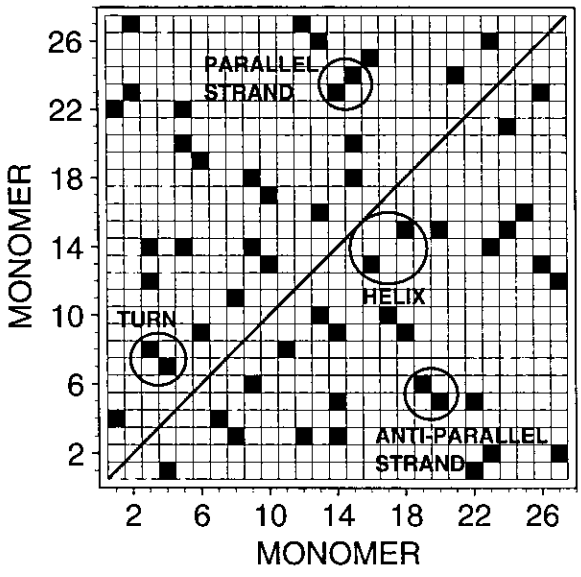

Figure 9. Definition of secondary structure segments (Chan & Dill, 1991). A contact map for sequence 43 is shown. A dark stippled square indicates a contact between the 2 corresponding monomers. The minimal length of a helix, parallel strand and anti-parallel strand is 2 contacts. A turn always has 2 contacts. Anti-parallel strand and helix contacts can include turn contacts.

expected, no significant differences in this preference can be seen between the three groups of structures.

Even though the structure of a 27-mer chain on a lattice is simple compared to that of real proteins, secondary structure can still be defined on the basis of characteristic contacts in the contact map (Chan & Dill, 1991; Fig. 9). Four classes can be distinguished: helix, parallel-sheet, antiparallel-sheet, and turn contacts. A helix, for example, is defined by at least two contacts of the form $i+\delta$, $i+2+\delta$, where i is the starting residue of the helix and δ has to change from 0 in steps of 2. There are no significant differences in the fractional content of the four secondary structure types between native structures of the folding and non-folding sequences (Table 1). As pointed out in the Discussion, the role of secondary structure in folding of real proteins and of longer chains on a lattice may be more important than observed here.

(f) *Association of foldicity with the correlation between interaction matrix and native structure*

A contact map **C** can be constructed for any conformation by placing 0 at position i,j if monomers i,j are in contact and 1 if they are not (Fig. 9). It has been shown that it is possible to produce successful folding by using an interaction matrix **B** that mirrors the contact map **C** for the native conformation; i.e. monomers that are in contact in the native state are made to interact attractively while other monomer pairs do not interact (Gō & Abe, 1981). This correlation means that a special bias towards the native state is introduced in the choice of parameters. Since random B_{ij} values are used in this study we can address the question whether a higher correlation between the interaction matrix **B** and the contact matrix **C** is a feature that distinguishes folding from non-folding sequences.

A plot of foldicity *versus* Pearson correlation coefficient between **B** and **C** is shown for the 200 sequences in the database (Fig. 10). Both folding and non-folding sequences have correlation coefficients of around 0·35. This relatively high correlation is due to validity of the "quasi-chemical approximation" according to which the probability of occurrence of a contact in the native state is proportional to $\exp(-B_{ij}/T_c)$ (Gutin *et al.*, 1992). However, there is no significant difference between the two groups in the sense that the variation within the two groups is smaller than the difference between the two groups (Fig. 10). Clearly, the correlation coefficients for the random folding and non-folding sequences are much closer to each other than to the correlation coefficient of unity for the sequences with perfect correspondence between **B** and **C**, as assumed in the Gō & Abe (1981) model. Sequences with a correlation coefficient between **B** and **C** significantly higher than the average do not occur among the 200 random sequences in the database. It follows that studying the folding of sequences where **B** is artificially highly correlated with **C** may lead to conclusions not generally valid for random folding sequences. Nevertheless, a small bias in the potential (eqn (1)) of the type described by Gō & Abe (1981) must be present to make the observed ground state most stable.

Whether or not the native structures of folding and non-folding sequences differ in their ability to

Table 1

Association between foldicity and secondary structure contents in the native conformation

Sequence type	Helix	Parallel sheet	Anti-parallel sheet	Turn
Strongly folding	0·118(±0·013)	0·416(±0·033)	0·264(±0·027)	0·121(0·020)
Non-folding	0·102(±0·006)	0·432(±0·014)	0·252(±0·011)	0·133(±0·009)

The fractional content of the contacts of different secondary structure types is shown. The anti-parallel sheet contacts include turn contacts. The number of all contacts is 28. The error is the standard error of the mean for the 30 strongly folding and 146 non-folding sequences.

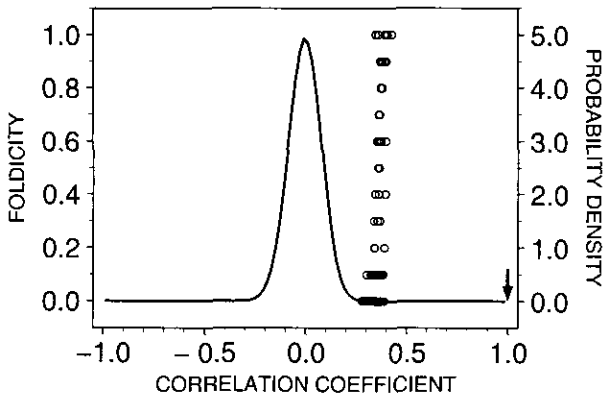

Figure 10. Dependence of folding on the correlation between the interaction matrix and the native contact map. This dependence is shown for the 200 random sequences in their native state (open circles). To judge the significance of the observed correlation coefficients, the curve shows the distribution of correlation coefficients for comparison of the 103,346 compact self-avoiding chains with random interaction matrices that were calculated using eqn (2) with $B_o = -2$ and $\sigma_B = 1$. The correlation coefficients for the interaction matrices used by Gō & Abe (1981) are 1, as indicated by the arrow.

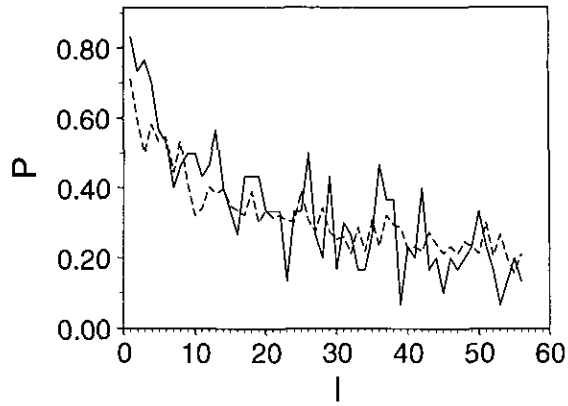

Figure 11. Probability that an l-th lowest B_{ij} will correspond to a native contact. The distribution was obtained as described in Results from the 200 random sequences. Strongly folding sequences (continuous line), non-folding sequences (broken line). The smaller noise in the curve for the non-folding sequences as compared with the folding sequences reflects the larger number of the non-folding sequences available to estimate the probability distribution.

satisfy the contacts with the lowest values of B_{ij} can be explored further as follows. The 156 B_{ij} interaction parameters in the **B** matrix that can possibly correspond to a contact between two monomers are sorted in order of increasing energy, separately for each of the 200 sequences. Every contact is then checked to determine whether it occurs in the native structure of the corresponding **B** matrix. Finally, a histogram is prepared that counts how many times the lowest B_{ij}, second lowest B_{ij}, third lowest B_{ij}, etc. corresponds to a native contact. This histogram is then normalized by division with the number of sequences to obtain the probabilities that the i-th lowest B_{ij} corresponds to a native contact (Fig. 11). As with the correlation coefficient between **B** and **C**, there is no difference between the folding and non-folding sequences. There is 80% chance that the lowest B_{ij} corresponds to a native contact. This probability decreases to 20% for the 50-th lowest B_{ij}. A fit of the curves $P[\langle B_{ij}(\Delta i)\rangle]$ in Figure 11 to $\exp[-\langle B_{ij}(\Delta i)\rangle/T_c]$ gives $T_c = 1{\cdot}7$. This value is significantly higher than $T_c \sim 1$ as estimated from $\langle X^{csa}(T)\rangle$ (Fig. 17(b), eqn (5)); it is also higher than $T_c \sim 0.77$ as estimated from the distribution of the energy gap between the two compact states with the lowest energies (Fig. 14; eqn (21)). This discrepancy reflects the degree to which the constraints imposed on the conformation by the chain connectivity invalidate the quasi-chemical approximation.

(g) *Association of foldicity with a relationship between energy levels and corresponding structures*

It was suggested by Shakhnovich *et al.* (1991) that the sequences with high foldicity have a significant optimum in $P_T^{csa}(Q)$ in the neighborhood of $Q = 0{\cdot}6$ to $0{\cdot}9$ in addition to the maximum at $Q = 1$; this optimum was absent in the non-folding sequences. The optimum is due to the existence of conformations that have low energies and are also structurally similar to the native state. The presence of these conformations was thought to speed up the folding kinetics.

Average $P_{T_x}^{csa}(Q)$ for the folding and non-folding sequences are shown in Figure 12. There are no significant differences between the folding and non-folding sequences. In particular, as demonstrated by the similarity of the two curves in the region $Q = 0{\cdot}6$ to $Q = 0{\cdot}9$, the non-folding sequences have the same distribution of the low energy conformations structurally similar to the native state as the folding

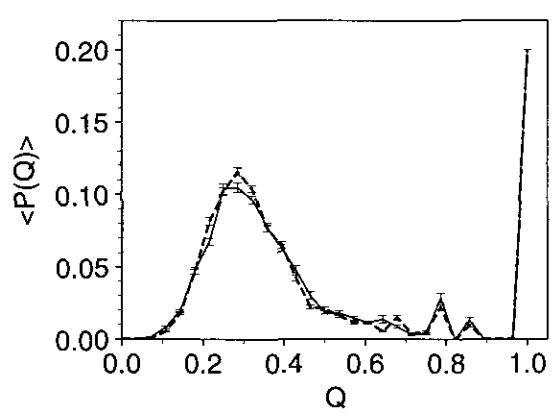

Figure 12. $\langle P_{T_x}^{csa}(Q)\rangle$ for strongly folding and non-folding sequences. The shape of $\langle P_{T_x}^{csa}(Q)\rangle$ for the folding sequences remains the same when only 15 sequences with foldicity of at least 0·80 are used in the average. Continuous and broken lines are used for the folding and non-folding sequences, respectively. The error bars indicate the standard error of the mean.

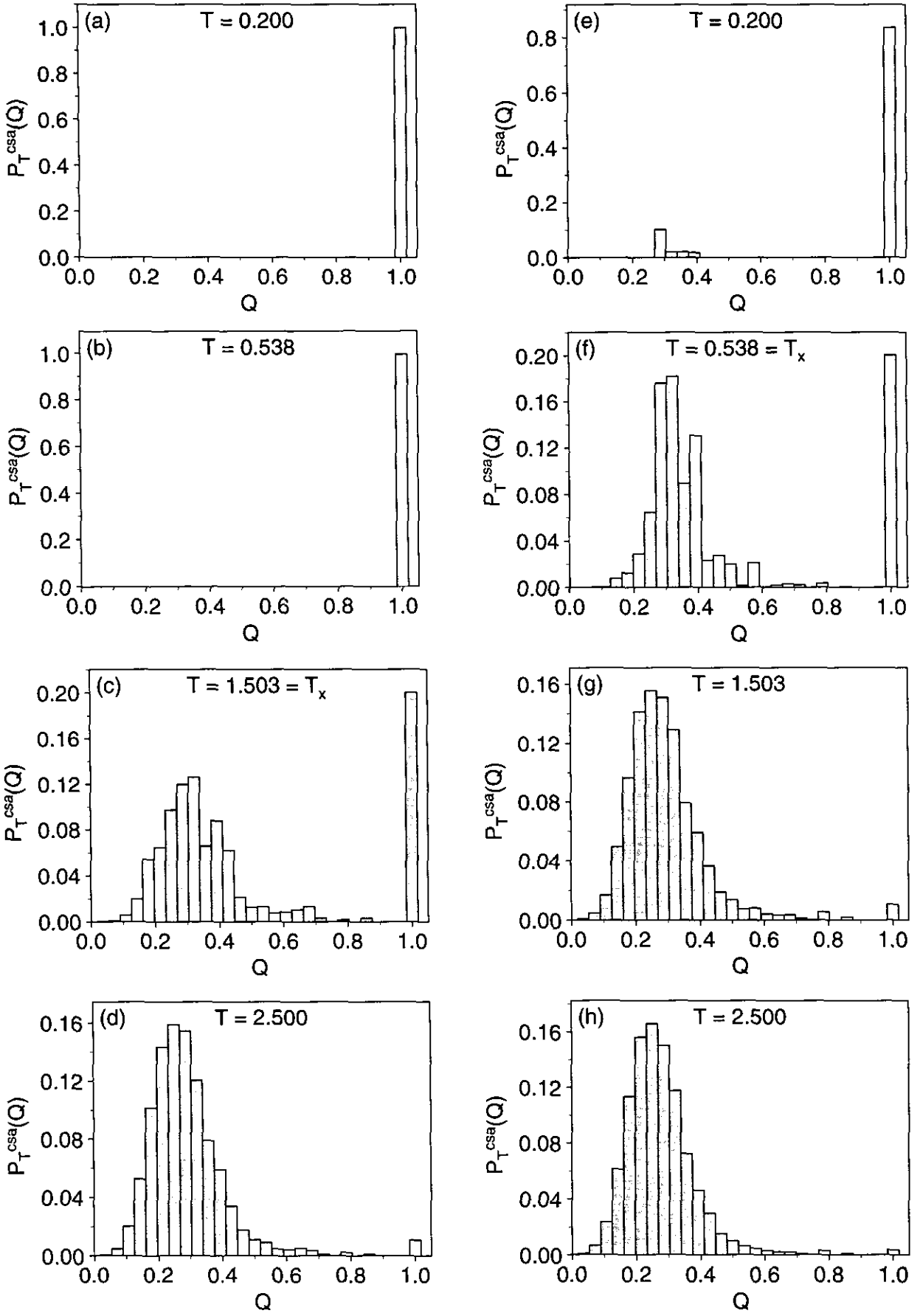

Figure 13. Comparison of $P_T^{csa}(Q)$ for folding sequence 43 and non-folding sequence 48. (a) to (d), folding sequence 43; (e) to (h), non-folding sequence 48. Temperatures T and T_x are indicated on the plots.

sequences. Because $T = T_x$ ($\tau = 1$) was used in Figure 12, the sequences are being compared at different values of T with T generally higher for the folding than the non-folding sequences. When $P_T^{csa}(Q)$ values are compared at the same absolute temperature T, the difference between the folding and non-folding sequences can be significant (Shakhnovich *et al.*, 1991; Fig. 13).

In principle, there are two possible reasons for a difference between $\langle P_T(Q)\rangle$ values for the two types of sequences. First, the lower part of the energy spectrum for the non-folding sequences might be compressed relative to the folding sequences while keeping the ordering of the energy levels intact. Second, the ordering of the levels might be changed by moving the low energy structures related to the native state to more positive energies; this would significantly decrease Boltzmann weights of these structures. Because this would change the form of $\langle P_{T_x}(Q)\rangle$ (e.g. it would decrease $\langle P_{T_x}(Q \sim 0{\cdot}8)\rangle$), it is clear from Figure 12 that the first case applies. In other words, the differences between the $\langle P_T(Q)\rangle$ values are due only to a difference in the folding temperature T_x and not to a different relationship between the energy levels and their structures. Note that T_x depends only on the distribution of energy levels, not on the corresponding structures, whereas $P_{T_x}(Q)$ does depend on the structures described by their monomer–monomer contacts. The lower T_x, the more compressed is the lower part of the energy spectrum. Therefore, the absolute temperature has to be decreased for the non-folding sequences to achieve the same thermodynamic preponderance of the native and near-native states with respect to the non-native states. The dependence of $P_T(Q)$ on T for a sample folding sequence and a non-folding sequence illustrates these statements (Fig. 13). It shows how the disappearance of a peak at $Q = 1$ for the non-folding sequence lags behind that of the folding sequence, corresponding to a lower T_x of the non-folding sequence. Additionally, it shows that the difference in the shape of $P(Q)$ between the non-folding and folding sequences can be accounted for by the temperature at which $P(Q)$ is calculated; i.e. the shapes in the following pairs of plots are similar: (b) and (e), (c) and (f), (d) and (g) (Fig. 13).

4. Discussion

The protein model used in this work is a 27-bead heteropolymer chain on a cubic lattice (Shakhnovich *et al.*, 1991). Despite its simplicity, the model contains essential features of protein folding. The total number of conformations, including the non-compact conformations, is approximately 5^{26}, a number much larger than can be scanned by any simulation. Thus, the Levinthal paradox is present in this system. The energy of each conformation is given by the sum of the monomer–monomer contact energies (eqn (1)); two monomers are in contact when they are nearest neighbors on the cubic lattice and are not neighbors in a sequence. The contact interactions were selected from a Gaussian distribution biased toward compact states by a mean (hydrophobic) attraction (eqn (2)). In this model, the density of levels in the energy spectrum of a sequence is also given by a Gaussian distribution (Appendix I). As a result, there is a discrete lower part of the energy spectrum with few levels and the quasi-continuous part with many levels that have more positive energies (Shakhnovich & Gutin, 1989, 1990*a*). The folding in this model is simulated by a Monte Carlo procedure that consists of a large number of small changes in the conformation, while preserving bond lengths and avoiding multiple occupancy of lattice sites.

Because the total number of compact self-avoiding conformations unrelated by symmetry is only 103,346 (Shakhnovich & Gutin, 1990*b*), the model permits the knowledge of the ground state energy and structure, as well as of the low energy spectrum with its associated structures. As one goes higher in energy towards the quasi-continuous range, the non-compact structures, which cannot be fully enumerated, become more important. However, this does not affect the results and the analysis proposed here.

We have generated and studied 200 random sequences of which 30 have high foldicity; i.e. in ten Monte Carlo runs at least four of the trajectories reach the ground state within 50×16^6 steps. Analysis of the results permits us to obtain information about what distinguishes the two types of sequences without a preconceived model. A significant aspect of this work is that the features important for folding are identified by comparison of the folding and non-folding sequences, not by inspection of the folding sequences alone, as in usual analyses of protein structures. Thus, this approach allows us to separate the properties exclusive to the folding sequences from the general features of any heteropolymer sequence. It leads to surprising results: some features which might have been expected to be typical of the folding sequences are also found in the non-folding sequences. First, we discuss only the lattice model and conclusions based on it. The question of the relation of this model to real proteins is considered in the second part of the Discussion.

Whether or not a given sequence folds rapidly to a stable native state is determined by features of its energy spectrum (Figs 5 to 7). There tends to be a large gap between the two lowest energy structures of the folding sequences. Moreover, a high $T_x = T(X^{csa}(T) = 0{\cdot}8)$ (eqn (10)) is conducive to high foldicity. $X^{csa}(T)$ (eqn (3)) is an order parameter that decreases with the thermodynamic weight of the native state. Thus, T_x increases with the weight of the native state at any given temperature. A high T_x results from a wide energy separation between the ground state and the bulk of the structures, which is proportional to the sparseness of the entire low energy spectrum, not only to the gap between the two lowest levels. This difference between the folding and non-folding sequences is manifested in the shape of the curve for $\langle X^{csa}(T)\rangle$ as a function of

T (Fig. 6). For the non-folding sequences, a non-co-operative curve, characteristic of the random sequences (Shakhnovich & Gutin, 1989; Fig. 17(b)), is obtained. By contrast, the folding sequences have a high co-operative (sigmoidal) shape for $\langle X^{csa}(T)\rangle$. It remains to be seen what patterns in the amino acid sequences give rise to the interaction matrix that results in the high T_x and in the large energy gap, and thus lead to rapid folding.

A number of proposals have been made in the literature concerning the mechanism by which a sequence can fold in a reasonable time, overcoming the problem raised by the Levinthal paradox. The present simulation permits us to address some of these. No difference exists between folding and non-folding sequences with respect to features of the native state, such as the distribution of the short and long range contacts (Fig. 8), and the contents of the helix, parallel sheet, anti-parallel sheet and turn contacts (Table 1). There is also no significant difference in the correlation between the interaction matrix **B** and the native contact map, although the correlation itself is significant for both the folding and non-folding sequences (Figs 10 and 11). We emphasize, however, that these findings were obtained for short chains on a lattice and that they may or may not be valid for longer chains of real proteins. To explore the differences in the relationship between the energy levels and corresponding structures, the probability distributions $P_{T_x}(Q)$ (eqn (9)) for the structural overlap Q were compared for the folding and non-folding sequences (Figs 12 and 13). The comparison shows that the lower part of the energy spectrum of the non-folding sequences with the associated structures can be obtained by compressing the lower part of the energy spectrum of the folding sequences. This leads to a smaller T_x. No reordering is required of the energy levels corresponding to the conformations that are close to the native state.

There are three temperatures of interest. The critical temperature T_c, which is a property of the ensemble of all random sequences (eqn (5)), the absolute temperature T, and the relative temperature τ, which is defined as the ratio between T and T_x (eqn (10)). Since $X^{csa}(T)$ varies with sequence, the relative temperature at the same absolute temperature is different for each sequence. At the same relative temperature τ, however, the sequences are likely to have approximately the same Boltzmann weight of the native state, regardless of the energy gap between the ground state and the first excited state (Fig. 3(a)). Whether a sequence folds is determined by an interplay between T and τ. The first condition is that absolute temperature T at which folding is attempted must be high enough (more than $\approx 1{\cdot}0\sigma_B/k_B$) so that the system has enough energy to overcome the highest energy barrier on the way to the native structure. The second condition is that relative temperature τ of folding must not be too high (i.e. must be less than $\approx 1{\cdot}0$), otherwise folding would not be sufficiently driven towards the global minimum and the native structure would not be stable enough. Instead, the sequence would be exploring too much of a phase space to fold in a reasonable time because its energy fluctuations would allow it to climb over more energy barriers. The relative temperature can be low enough when the lower discrete part of the spectrum is sparse, with the energy gap between the lowest and second lowest energy levels, ΔE_{10}, greater than $k_B T$. Therefore, the requirement for a folding sequence is that the relative temperature is sufficiently low at a sufficiently high absolute temperature. Both conditions are satisfied when T_x is larger than ~ 1, as shown in Figure 7.

As noted before, the only conditions for rapid folding into the stable native state that are revealed in this study are the large gap in the energy spectrum between the native state and other conformations, and the sparseness of the discrete part of the spectrum. This coincides with fulfilling the stability requirement at a high absolute temperature. In particular, the temperature of folding at which the native state is stable (T_x) has to be higher than the critical temperature for the ensemble of random sequences (T_c). In our model, $\langle T_x\rangle = 1{\cdot}24(\pm 0{\cdot}035)$ for the folding sequences, and $T_c \approx 1$ for the ensemble of all random sequences with a compact self-avoiding conformation (Fig. 17(b), eqn (5)). By contrast, $\langle T_x\rangle$ for the non-folding sequences is $0{\cdot}63(\pm 0{\cdot}016)$. Thus, the main difference between the folding sequences and the non-folding sequences is that the folding sequences are stable at a higher temperature where folding is kinetically possible.

A random folding sequence is different from random sequences (e.g. Fig. 6) because it satisfies the two folding conditions; i.e. it has a thermodynamically stable native state and this native state is kinetically accessible from the denatured state.

The thermodynamic requirement has been discussed by Shakhnovich & Gutin (1990*a*), who have derived an expression that gives the fraction of all random heteropolymer sequences with a specified Boltzmann weight of the native state at a given temperature (eqn (7)). This fraction is determined on the basis of the whole energy spectrum by taking into account the difference between the energies of the lowest state and all excited states, not only the first excited state. Equation (7) is applicable only below T_c and cannot be employed to estimate the fraction of thermodynamically stable sequences above T_c. Over the whole temperature range the fraction of thermodynamically stable sequences decreases as the temperature and the required probability of the native state increase. At each temperature, below and above T_c, the thermodynamically stable random sequences are different from the random sequences because they have a sufficiently large energy gap between the lowest state and the rest of the states (Shakhnovich & Gutin, 1990*a*).

The Monte Carlo simulations show that a random sequence is not likely to fold at a temperature where it is reasonably stable if that temperature is below

T_c (Fig. 7); T_c is approximately unity for the ensemble of compact self-avoiding chains (Fig. 17(b)). On the other hand, a random sequence is likely to be able to fold if it is reasonably stable at a temperature higher than T_c (Fig. 7). Random folding sequences are those thermodynamically stable random sequences that satisfy the thermodynamic requirement at a temperature higher than T_c; i.e. all random sequences that are thermodynamically stable at $T > T_c$ and none of the random sequences that are only stable at $T < T_c$ are likely to be folding sequences. Further discussion of this point will be given in a subsequent paper (A. Šali, E. Shakhnovich & M. Karplus, unpublished results).

The probability that a gap between the ground state and the first excited state is equal to or greater than Δ can be estimated from heteropolymer theory (Appendix I):

$$F(\Delta E_{10} > \Delta) \approx \exp(-\Delta/T_c). \qquad (12)$$

Since real folding (Gutin *et al.*, 1992) and simulations take place at $T \approx T_c$, we see that $\approx 1\%$ of all sequences have a gap of a few $k_B T$. In this paper, we require a native state concentration of only 40% (Fig. 3(a)) to be able to simulate folding at a temperature at which it is most rapid (Fig. 2). In this case, 15% (30/200) of the random sequences are folding sequences (Fig. 4). Note that we calculate thermodynamic stability using the compact self-avoiding chains only. Since there are many non-compact self-avoiding chains, the true thermodynamic stability is smaller than 40% as obtained with the use of the compact self-avoiding ensemble alone. Based on an exhaustive Monte Carlo sampling beyond the first passage time, the true thermodynamic stability of the folding sequences at T_x is estimated to be 1 to 5% (A. Šali, E. Shakhnovich & M. Karplus, unpublished results). At the folding temperature, the interaction energies of the native contacts are not sufficiently favorable for the chain to remain in the native state for a large fraction of the time. On the average, the more stable a structure as based on the maximally compact states, the higher its true stability. Consequently, if a more stable native state were required, the only difference would be that many more sequences would have to be simulated for a longer time at a lower temperature (eqn (12), Fig. 2), which is not computationally feasible. This situation can be compared with studies of long chains on the diamond lattice (Skolnick & Kolinski, 1991) where true thermodynamic stability was achieved by explicitly decreasing the energy of the native contacts. Another example is a study of folding transitions by the use of Monte Carlo sampling of heteropolymer chains on the 3D cubic lattice (O'Toole & Panagiotopoulos, 1992). In this study, the thermodynamic stability of the native state was ensured by designing optimal sequences for highly symmetric native structures. In contrast to these two studies, our purpose was not to *design* special stable sequences but to find out by selection from *random* sequences what is required for folding.

The essential feature of the interaction matrix **B** that results in a pronounced energy minimum is its correlation with the contact matrix of the native state. However, it is important to note that only a weak correlation is required. In fact, as shown here, the difference in the Pearson correlation coefficient between **B** and **C** for the folding and non-folding sequences is not significant in the sense that the variation within the two groups is smaller than the difference between the two groups. To illustrate the sensitivity of ΔE_{10} to the correlation between **B** and **C**, we consider the effect of making the native contact energies B_{ij} more favorable by only $0{\cdot}1\ k_B T$. Such a change to the native contacts increases ΔE_{10} by about $2{\cdot}8\ k_B T$. This increase in ΔE_{10} is comparable to the initial ΔE_{10} of the folding sequences (Fig. 7). Such a modification of **B** increases the cross-correlation coefficient by only 0·03. This small change in the cross-correlation coefficient is well within the variability observed among the random folding and non-folding sequences (Fig. 10).

Although it has been sufficient to consider only the energy spectrum in separating folding from non-folding sequences in our simulations, a simple argument shows that the requirement for rapid folding found in this work is not the only necessary feature of the energy hypersurface. We consider a typical folding sequence with a large energy gap and randomly shuffle the structures assigned to each of the energy levels. The resulting energy hypersurface could correspond to the golf course model of a rugged energy landscape (Bryngelson & Wolynes, 1989). This golf course hypersurface still satisfies the folding requirement as described in this work because the density of states is not changed by randomizing the structures. For such an energy hypersurface, however, no mechanism short of enumerating all possible conformations will find the native state. Thus, rapid folding on this surface does not occur. This argument seemingly contradicts the results in the present paper which show that a pronounced energy minimum is necessary and sufficient for rapid folding. However, this apparent contradiction can be explained in the following way. First, a physically reasonable polypeptide energy function does not result in the golf course landscape. This is so because of the chain connectivity and because a small change in the structure is likely to produce a small change in the energy. The latter effect is intrinsic to any energy function expressed as a sum of many small terms (for example, eqn (1)). The resulting non-randomness in the distribution of the structures among the energy levels is present in the folding sequences as well as in the non-folding sequences. Second, the simulations show that only those sequences with a pronounced minimum fold rapidly and that all the sequences with a pronounced minimum fold rapidly. This shows that either (1) no special distribution of structures among the different energy levels exists for the folding sequences as compared with the non-folding

sequences, or that (2) the additional non-randomness in the distribution of structures that results in rapid folding is created with high probability when there is a large energy gap between the lowest two states and when the lower discrete part of the spectrum is sparse. The former possibility is supported by the plots of $\langle P^{csa}(Q)\rangle$, which show that no reordering of the structures assigned to the different energy levels is required to match the distributions of the folding and non-folding sequences (Fig. 12). A more complete description of how a large energy gap creates the conditions for rapid folding will be given in a separate paper (A. Šali, E. Shakhnovich & M. Karplus, unpublished results).

Two questions arise about the generality of the results obtained for the 27-mer lattice model. Are these results valid for longer polymeric chains? Are they valid for real proteins?

To give a complete answer to the first question, a simulation of significantly longer chains would have to be done. Such a simulation would be impossible to combine with exhaustive enumeration. However, the following facts indicate that the folding requirements described here may be valid for longer chains. Comparison of the thermodynamic results from simulation of the 27-mer with predictions of analytical theory for long chains (Shakhnovich & Gutin, 1989) shows no significant differences. Also, the time course of the process of folding (Shakhnovich *et al.*, 1991; A. Šali, E. Shakhnovich & M. Karplus, unpublished results) is consistent with the thermodynamic description of random polymers given in the analytical studies of long chains; i.e. the folding transition is a cooperative first order phase transition with a significant free energy barrier between the denatured and native states. It has been shown that the probability distribution for a gap size, ΔE_{10}, in the energy spectrum does not depend on chain length (eqn (12)). Further, the results obtained do not depend on the details of the Monte Carlo folding algorithm nor on the details of the energy function; e.g. when several monomers are allowed to occupy the same lattice point with some energetic penalty and the crankshaft moves are not performed, the associations between foldicity and polymer features do not change (Appendix II). Moreover, Monte Carlo folding simulations with random 16-mer and 25-mer sequences on the two-dimensional cubic lattice result in the same conclusions (A. Dinner, A. Šali, E. Shakhnovich & M. Karplus, unpublished results). The time scale of folding of real proteins varies widely but does not appear to be dependent on the chain length (Roder & Elöve, 1993). It should be noted, however, that longer chains may lead to a greater importance of secondary structure in the kinetics of the folding process.

Other simulations based on lattice models of polypeptide chain that were constructed to obtain rapid folding (Gō & Abe, 1981; Leopold *et al.*, 1992) have large energy gaps on the order of Nk_BT energy units. In the model described by Gō & Abe (1981), only native contacts have a non-zero contact energy on the order of $-k_BT$. As a result, the negative energy of a conformation is proportional to its fraction of native contacts. Because chain connectivity severely limits the number of conformations with a significant fraction of native contacts, the gap between the native and the few near-native conformations on one hand, and the bulk of the remaining conformations on the other hand, is large, in the order of Nk_BT. No structure that is significantly different from the native state can have low energy. Similar reasoning applies to the model described by Leopold & co-workers (Leopold *et al.*, 1992) since their interaction matrix is related to that of Gō & Abe. Our results are also in agreement with the assumption of Goldstein *et al.* (1992) that folding competes with the glass transition and that folding is maximized when the ratio between the folding and the glass transition temperatures is maximized. They used the REM model to show that this ratio increases with the energy difference between the native state and the "liquid-like" phase and decreases with the energy fluctuations in the liquid-like phase. Since the liquid-phase in their model corresponds to the quasi-continuous part of the spectrum, their assumed criterion is similar to the requirements for a large energy gap and for the sparseness of the lower part of the spectrum that were found in this work. Recently, Miller *et al.* (1992) studied folding of short heteropolymers with hydrophobic and hydrophilic beads by a Monte Carlo simulation on the 2D square lattice. They found that folding is slow when T is too low because the chain becomes stuck in local minima, that folding is slow when T is too high because the molecule searches randomly through the large ensemble of open conformations, and that folding may be relatively fast for intermediate T. These findings are consistent with our results.

A more complex question is whether the present results are valid for real proteins. One important difference is that the lattice model does not take into account the presence of side-chains. Those are known to be tightly packed in the native state (Ponder & Richards, 1987) and their packing may be involved in the final stage of protein folding (Ptitsyn, 1987; Shakhnovich & Finkelstein, 1989). Thus, the correspondence between the folding of the lattice model and real proteins is likely to be as follows. The first process in both cases is a fast transition from a random coil with low density into a random homopolymer-like state with a relatively high density and random structure (Elöve *et al.*, 1992; Roder & Elöve, 1993; Shakhnovich *et al.*, 1991). Many proteins then fold into the native state through another intermediate, the molten globule (Ptitsyn, 1987), which has a high density and a structure similar to the native state but may differ from it in the details of the side-chain and secondary structure packing. In the lattice model, the folding proceeds directly from the high density non-native state to the unique state which is equated with the native state (Shakhnovich *et al.*, 1991). Thus, it may be better to regard the present

model as concerned with the transition from a random coil to a collapsed globule and finally to the organized molten globule. For real proteins, during the conversion of the molten globule into the native structure, specific stereochemical requirements have to be fulfilled to allow tight packing. It has been argued that such requirements can be satisfied with significant probability even for random sequences (Ptitsyn & Volkenstein, 1986).

It seems likely that real proteins also have a Gaussian energy density because the energy is a sum of a large number of weakly correlated energy terms. Moreover, some interdependence between the energies of states with similar conformations is expected for the same reasons that were given for the energy function in equation (1). In such a case, the sufficient requirement for rapid folding described in this paper, the sparseness of the lower part of the spectrum guaranteed for a small fraction of sequences by the Gaussian distribution of energy levels, would be directly applicable to real proteins even though the exact structural model and energy function were not.

Another difference to be considered when applying the present lattice model to proteins is that protein motion is guided by molecular dynamics whereas the lattice model folds according to the Monte Carlo procedure. Monte Carlo simulation is an approximation to molecular dynamics because it allows only discrete particle positions, a small number of moves that transform a system from one discrete configuration to another, and because it assigns arbitrary rates to these moves. There is some evidence that the main features of the Monte Carlo folding process do not depend on whether the diamond lattice or the 210 cubic lattice are used (Skolnick & Kolinski, 1991). Moreover, as discussed above, no differences in the important features of folding kinetics have been observed between the 2D (A. Dinner, A. Šali, E. Shakhnovich & M. Karplus, unpublished results) and the two 3D lattice models employed in this paper; the moves allowed in the 2D model were of the same type as the moves in the present 3D model. These observations suggest that Monte Carlo procedure is adequate for describing the main features of the protein folding process, especially when those features occur over many Monte Carlo steps. Ultimately, this assumption can only be tested by comparing the current results with those from a molecular dynamics study, which is currently very difficult to do even for a model as simple as the one used here. Tests could be made for 2D models, though they would require some way of estimating the probability of the different pseudo-angle and pseudo-dihedral angle transitions that correspond to the Monte Carlo moves. The equivalency of Monte Carlo simulations and molecular dynamics is consistent with a separation between the deterministic small scale and stochastic large scale motions on short and long time scales, respectively. Positive conclusions on the applicability of the Monte Carlo method to study dynamics of proteins were also obtained by Skolnick & Kolinski (1991). We note that explicit inclusion of the rigid body Monte Carlo moves of larger elements of secondary structure is useful in simulations of more complex protein-like models (Skolnick & Kolinski, 1991). However, we do not include these types of moves in the present simulations because even the local moves of one or two monomers are sufficient for a significant fraction of random sequences to be folding sequences and because our simulations are already ergodic (A. Šali, E. Shakhnovich & M. Karplus, unpublished results).

The folding simulations show that rapid folding always ends in the global energy minimum and that random sequences do not repetitively fold into the same local minimum. This indicates that the native state corresponds to the global energy minimum, and not to one of the local minima as suggested by the metastability hypothesis of the native state (Honeycutt & Thirumalai, 1992). If this finding is valid for real proteins, it implies that protein structure prediction methods may not have to simulate the whole folding process but could rely on energy minimization methods. It is conceivable that these methods could take a more direct route through the phase space when searching for the global energy minimum, although the multi-minimum problem still has to be solved (Scheraga, 1989).

Recently, Rooman *et al.* (1992) used a database of known protein structures to derive pseudo-energy potentials for main-chain conformation of a single residue. They applied the potentials to predict the structure of segments between 5 and 15 residues long. They showed that those segments that have large energy gaps between the two lowest states are predicted most reliably. Moreover, in many cases the same regions were identified as the parts of the protein structure that are likely to form early in protein folding (Rooman & Wodak, 1992). This result is consistent with our simulations that indicate that the whole sequence with a large gap folds fast, from which it is reasonable to expect that the parts of the whole sequence also have the same tendency during early stages of folding when short segments independently assume non-random structure.

One interesting extrapolation of the importance of a pronounced energy minimum as the necessary and the sufficient condition for rapid folding is its use to propose methods for design of stable folding proteins (Shakhnovich & Gutin, 1993*a*,*b*).

In its more general form, the problem addressed in this paper is: what are the characteristics of the many-dimensional surface of a non-linear function with many interdependent local optima that will allow a Monte Carlo procedure to find a global optimum? The answer is that a Monte Carlo procedure tends to be successful if the function has a pronounced global optimum. This means that the success of optimization can be improved not only by better optimizers but also by devising functions with a more pronounced global optimum. This result was used as an assumption to help in the design of associative-memory Hamiltonians for

recognizing protein folds (Goldstein *et al.*, 1992). Other examples of optimizing methods that take advantage of a more pronounced optimum by simplifying the function include the variable target function method (Braun & Gō, 1985), the antlion method (Head-Gordon & Stillinger, 1993) and the diffusion equation method (Piela *et al.*, 1989). The global optimum can also be made more pronounced by using additional information as exemplified by the methods relying on NMR derived constraints (Clore *et al.*, 1986), X-ray diffraction data (Brünger *et al.*, 1987) and related protein structures (Šali & Blundell, 1993).

APPENDIX I

Statistics of the Lowest Energy Gap in Random Sequences

To examine the statistics of the size of the energy gap ΔE_{10} between the ground-state and the first excited state, we use the Random Energy Model (REM) developed for spin glasses (Derrida, 1981). This model has been shown to describe the thermodynamics of heteropolymers (see the Discussion for applicability to proteins and the definition of the REM in eqn (2): Shakhnovich & Gutin, 1989).

In the REM, the energies of the conformations are independent random variables distributed according to the following Gaussian density function (Shakhnovich & Gutin, 1989):

$$P_E(E) = \frac{1}{\sqrt{2\pi}\sigma_E} e^{-\frac{1}{2}\left(\frac{E-E_o}{\sigma_E}\right)^2},$$

$$\sigma_E = \sigma_B\sqrt{N\rho}, \quad E_o = N\rho B_o, \quad (13)$$

where ρ is the average number of contacts per monomer; each contact is counted only once (eqn (1)). The expression for the fraction of sequences having the lowest gap larger than Δ is:

$$F(\Delta E_{10} > \Delta) = M\int_{-\infty}^{\infty} P(E)\left[\int_{E+\Delta}^{\infty} P(\xi)\mathrm{d}\xi\right]^{M-1}\mathrm{d}E, \quad (14)$$

where $P(E)$ is given by equation (13) and $M = \gamma^{N-1}$ is the total number of conformations. Equation (14) states that each of the M conformations may be a ground state conformation with energy E and that the remaining $M-1$ conformations must have an energy at least $E+\Delta$ each.

Defining:

$$V(\xi) = 1 - \int_{-\infty}^{\xi} P(v)\mathrm{d}v, \quad (15)$$

and making a substitution $E+\Delta = U$ we obtain:

$$F(\Delta E_{10} > \Delta) = -M\int \left.\frac{\mathrm{d}V}{\mathrm{d}U}\right|_{U-\Delta} V^{M-1}(U)\mathrm{d}U. \quad (16)$$

We used the relation:

$$\frac{\mathrm{d}V}{\mathrm{d}U} = -P(U),$$

which follows from the definition of V (eqn (15)). For $P(E)$ of a Gaussian form (eqn (13)) we have:

$$\left.\frac{\mathrm{d}V}{\mathrm{d}U}\right|_{u\ \Delta} = \frac{1}{\sqrt{2\pi}\sigma_E} e^{\left(\frac{U-E_o-\Delta}{\sigma_E}\right)^2}$$

$$= P(U)\, e^{\left(\frac{2(U-E_o)\Delta}{\sigma_E}\right)} e^{-\left(\frac{\Delta}{\sigma_E}\right)^2}. \quad (17)$$

Later we will see that the largest contribution to the integral in equation (16) comes from:

$$|U_c| \sim \sigma_E/\sqrt{N} \sim N$$

and $\Delta \sim 1$; therefore the last term in the exponential in equation (17) may be omitted and we get:

$$F(\Delta E_{10} > \Delta) = -M\int \left.\frac{\mathrm{d}V}{\mathrm{d}U}\right|_U V^{M-1}(U) e^{\frac{2U\Delta}{\sigma_E^2}}\mathrm{d}U$$

$$= -\int \frac{\mathrm{d}(V^M)}{\mathrm{d}U} e^{\frac{2U\Delta}{\sigma_E^2}}\mathrm{d}U. \quad (18)$$

Integrating by parts and taking into account that $V(\infty) = 0$ and $V(-\infty) = 1$ we have:

$$F(\Delta E_{10} > \Delta) = \frac{2\Delta}{\sigma_E^2}\int V^M(U) e^{\frac{2U\Delta}{\sigma_E^2}}\mathrm{d}U$$

$$= \frac{2\Delta}{\sigma_E^2}\int e^{-M\int_{\infty}^{U} P(\xi)\mathrm{d}\xi + \frac{2U\Delta}{\sigma_E^2}}\mathrm{d}U, \quad (19)$$

where we used the definition of V (eqn (9)) and the fact that $M \gg 1$. The integrand in equation (19) vanishes when $U > E_c$ where E_c is determined from the condition:

$$M\int_{-\infty}^{E_c} P(\xi)\mathrm{d}\xi = 1. \quad (20)$$

Taking P as given in equation (13) we have for

$$E_c = -\sqrt{N}\sigma_E\sqrt{2\rho\ln\gamma},$$

which coincides with the boundary between the semi-continuous and lower discrete parts of the spectrum in the REM (Shakhnovich & Gutin, 1990*a*). We approximate the integral in equation

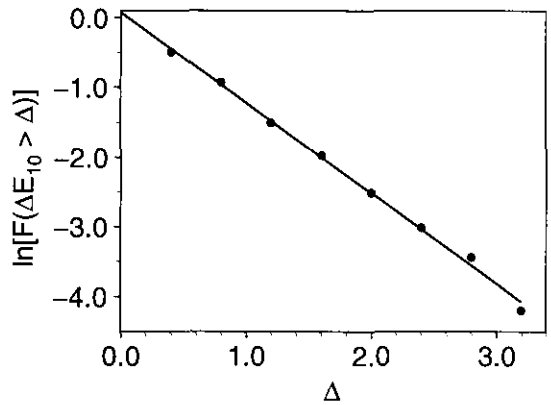

Figure 14. Comparison of $\ln[F(\Delta E_{10} > \Delta)]$ with the results from a simulation. A total of 1000 random sequences were generated using eqn (2) to obtain the experimental distribution (points). The line is a least-squares fit of the analytical model in eqn (12) to these points. The best value for parameter T_c was found to be 0·77.

(19) and finally get:

$$F(\Delta E_{10} > \Delta) = \frac{2\Delta}{\sigma_E^2}\int_{-\infty}^{E_c} e^{\frac{2U\Delta}{\sigma_E^2}} dU = e^{-\frac{E_c\Delta}{\sigma_E^2}} = e^{-\frac{\Delta}{T_c}}. \quad (21)$$

Note that $F(\Delta E_{10} > \Delta)$ does not depend on the length of a chain even though the density of energy levels, $P(E)$, does (eqn (13)).

The distribution $F(\Delta E_{10} > \Delta)$ depends on the parameter T_c. This parameter was estimated by least-squares fitting the model in equation (21) to the distribution of ΔE_{10} values for 1000 random sequences (Fig. 14). $T_c = 0{\cdot}77$ was found to give a good fit to the model. This T_c value for the ensemble of compact self-avoiding chains can be compared with $\langle X^{csa}(T)\rangle$ using equation (5). $\langle X^{csa}(T)\rangle$ is linear for $T < \approx 0{\cdot}77$, thus confirming the applicability of the model in equation (21). T_c from equation (21) can also be compared with an independent calculation based on equation (6). For the ensemble of self-avoiding chains, $\gamma = 4{\cdot}17$ from $M^{csa} = (\gamma/e)^N$ where $M^{csa} = 103{,}346$. The average number of contacts per monomer in the ensemble of compact self-avoiding chains is 28/27, thus giving $T_c = 0{\cdot}85$ (eqn (6)). This value is in agreement with $T_c = 0{\cdot}77$ from equation (21). Although T_c for the ensemble of all self-avoiding chains is not estimated here, it must be lower than T_c for the compact self-avoiding chains because more structures occur in the discrete part of the spectrum (Fig. 17).

APPENDIX II

Dependence of Folding on a Monte Carlo Folding Algorithm

Shakhnovich *et al.* (1991) did a similar study for the 27-mer cube with a Monte Carlo algorithm that is slightly different from the one employed in this paper. Their algorithm allowed up to three monomers at the same lattice site and the energy function had added terms that penalized the double and triple occupancies. No crankshaft moves were included in the original model. However, the original model still led to rapid folding because multiple occupancies were allowed. A possible negative consequence of the multiple occupancies is that chain cutting may occur, which could change the characteristics of the folding model compared to the real folding process.

The 200 random sequences studied here were tested with the original Monte Carlo folding algorithm. To get the foldicities with this algorithm, up to 20×10^6 steps were done with original values for the multiple occupancy penalties of $D_2 = 10$ and $D_3 = 14$. The rest of the parameters were the same as for the current Monte Carlo algorithm.

A high correlation between the two types of foldicities is shown for the 200 random sequences in Figure 15. This indicates that the conclusions about the features important for rapid folding are independent of the details of the folding model used.

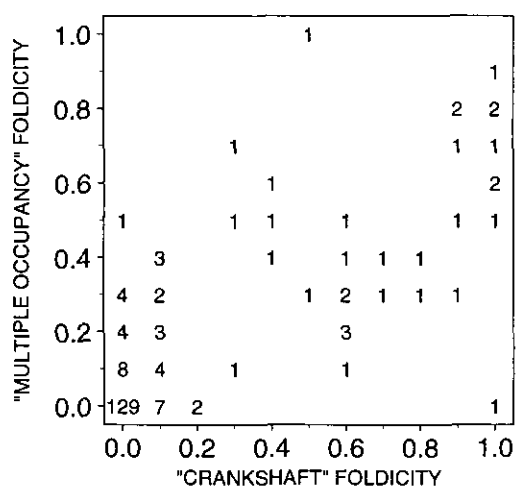

Figure 15. Comparison of foldicities from the 2 different Monte Carlo procedures. The numbers show how many sequences out of the sample of 200 have a corresponding combination of foldicities from the two Monte Carlo procedures. "Crankshaft" foldicity is obtained by the procedure described in Methods. "Multiple occupancy" foldicity is obtained by the procedure published by Shakhnovich *et al.* (1991) and described in Appendix II.

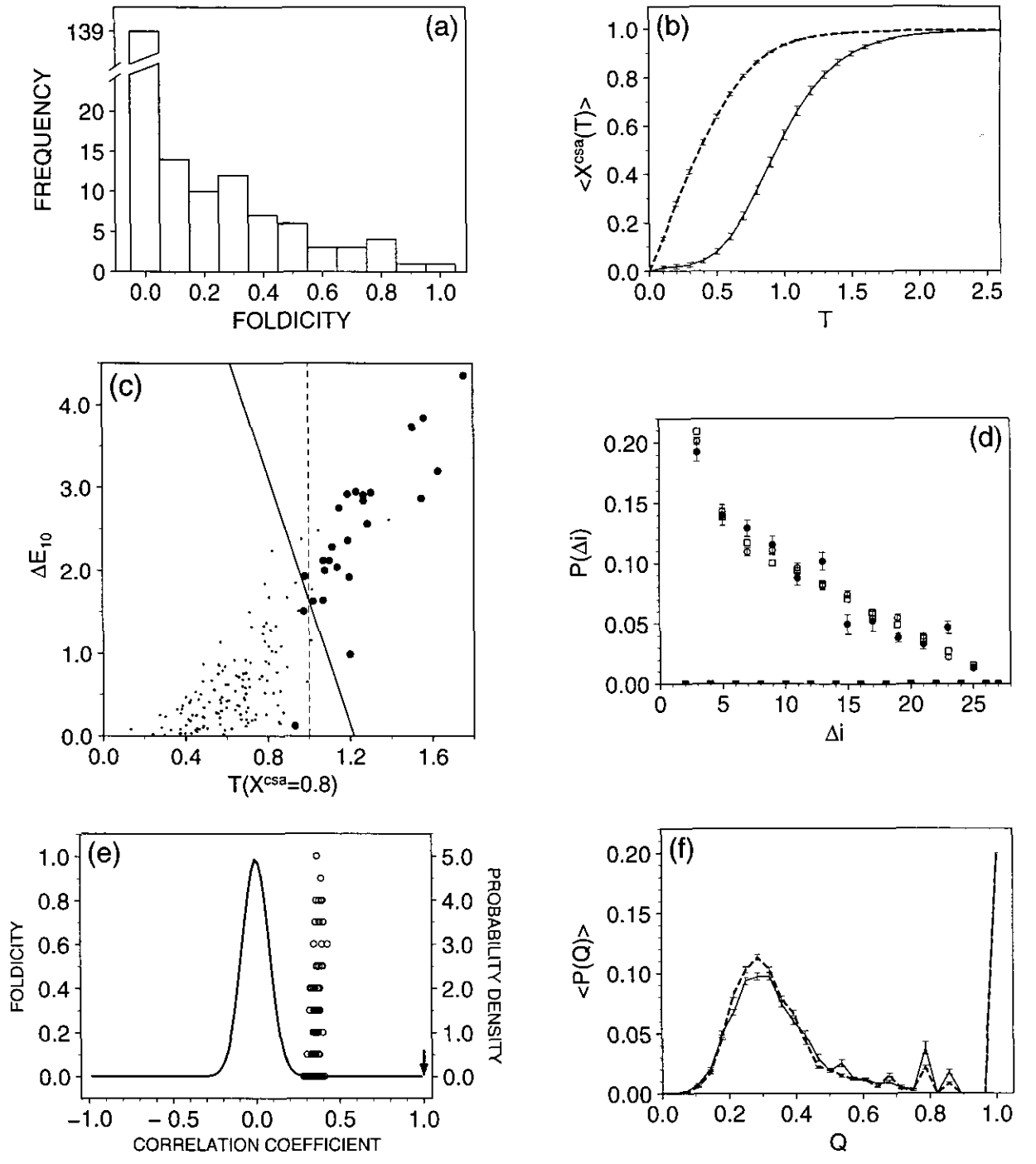

Figure 16. Associations for the multiple occupancy Monte Carlo algorithm. This Figure repeats the plots in Figs 4 (a), 6 (b), 7 (c), 8 (d), 10 (e) and 12 (f) with the data from the multiple occupancy Monte Carlo algorithm. See the legends to the original Figures for explanation. There are 25 strongly folding and 139 non-folding sequences for this algorithm.

Table 2

Association between foldicity and secondary structure contents in the native conformation for the multiple occupancy Monte Carlo algorithm

Sequence type	Helix	Parallel sheet	Anti-parallel sheet	Turn
Strongly folding	0·101(±0·013)	0·386(±0·036)	0·290(±0·029)	0·097(±0·018)
Non-folding	0·095(±0·006)	0·428(±0·014)	0·256(±0·011)	0·134(±0·009)

This Table is equivalent to Table 1, except that the 25 strongly folding and 138 non-folding sequences were defined by the multiple occupancy Monte Carlo algorithm.

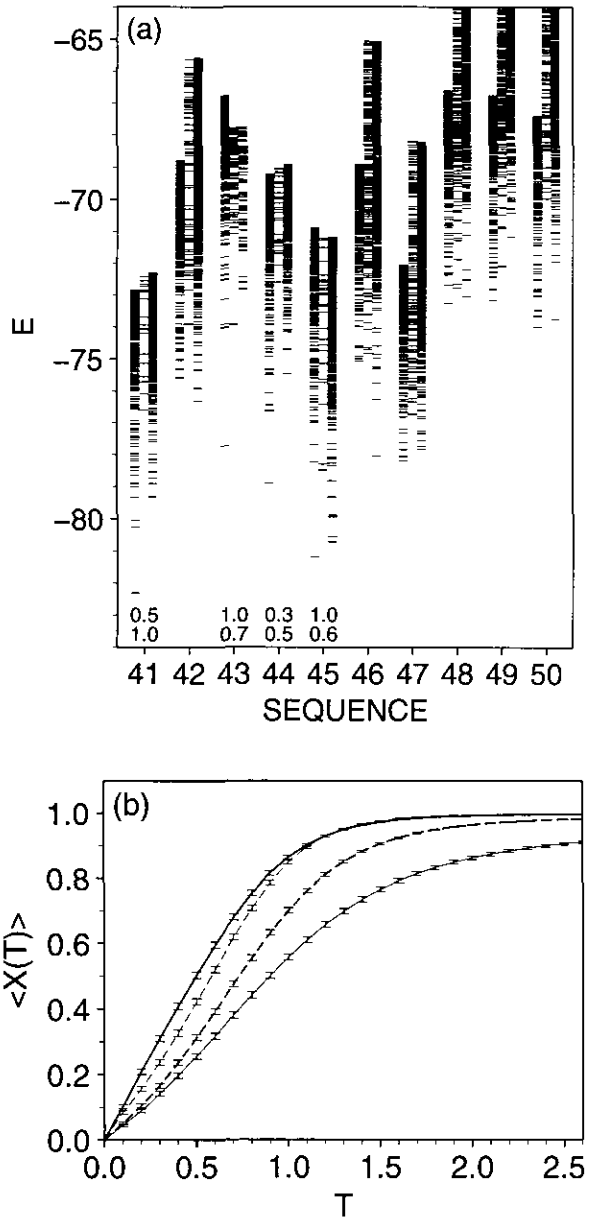

Figure 17. Comparison of the ensemble of compact self-avoiding chains with the ensembles that also contain non-compact self-avoiding chains and conformations with double and triple occupancies. (a) Comparison of the lower part of the energy spectra for sequences 41 to 50. The energy levels are shown for the lowest 400 compact self-avoiding chains (the 1st column), non-compact self-avoiding chains only (the 2nd column), and only the structures with multiple-occupancies (the 3rd column). Non-compact and multiple occupancy structures with $E < E_o + 10$ were sampled in a long multiple occupancy Monte Carlo simulation of 50×10^6 steps, starting from the native structure at a high temperature of $T = 1{\cdot}2$. The top and bottom line of numbers below the spectra are the crankshaft and multiple occupancy foldicities, respectively. (b) Comparison of $\langle X(T) \rangle$ for the 200 random sequences. Shown are $\langle X(T) \rangle$ for all compact self-avoiding structures (thick, continuous line), all compact self-avoiding structures from the Monte Carlo simulation (thin, continuous line), all self-avoiding structures from the Monte Carlo simulation (thick, broken line), and all structures from the Monte Carlo simulation (thin, broken line). The conformations for the calculation of the individual $X(T)$ curves in the last 3 averages were obtained from the Monte Carlo simulations described above. Eqn (3) with appropriate M was used in the calculation of each $X(T)$. The error bars are the standard errors of the mean.

The results from the multiple occupancy simulation were then analyzed in the same way as the results from the Monte Carlo simulation without multiple occupancies; Table 1 and Figures 2, 4, 6, 7, 8, 10 and 12 were replotted with the data from the multiple occupancy simulation (Table 2, Fig. 16). No significant differences were observed between the results from the two sets of simulations.

However, there are some differences between the two models. For the original algorithm, the discrete part of the spectrum contains many multiple occupancy structures when the energy parameters B_o, σ_B, D_2 and D_3 are optimized for rapid folding (Fig. 17(a)). Moreover, for 25% of the sequences, a multiple occupancy super-compact structure was found by a Monte Carlo simulation at high temperature that had a lower energy than the compact self-avoiding conformation with the lowest energy (i.e. the native state). For the ensemble with no multiple occupancies, only 5% of the sequences have a semi-compact conformation that is lower in energy than the native state. Consequently, the thermodynamic functions that depend on the discrete part of the spectrum are not identical for the ensemble of self-avoiding chains and the ensemble that also includes the chains with double and triple occupancies (Fig. 17(b)).

An important advantage of the multiple occupancy algorithm is its speed. The multiple occupancy folding simulation needs approximately 24 seconds of CPU time on an IBM RS/6000-550 workstation for 1 million Monte Carlo steps. The current algorithm spends about 30 seconds for the same number of steps but it also needs to do about three times as many steps to fold a sequence. This makes the multiple occupancy algorithm about four times more efficient than the current algorithm. Given the high similarity in the kinetic behavior of the two algorithms (Fig. 15), the multiple occupancy algorithm may be preferred for kinetic studies of protein folding.

We thank Aaron Dinner, Georgios Archontis and Peter Leopold for discussing the protein folding problem. E.S. benefited from numerous discussions with Alexander Gutin, Oleg Ptitsyn, Peter Wolynes and Alexei Finkelstein. A.Š. is 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 (M.K.). The computations were done on IBM RS/6000, Silicon Graphics Iris 4D, SUN Sparcstation, DEC Decstation and NeXT workstations.

References

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Braun, W. & Gō, N. (1985). Calculation of protein conformations by proton–proton distance constraints: a new efficient algorithm. *J. Mol. Biol.* **186**, 611–626.

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

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). Crystallographic *R*-factor refinement by molecular dynamics. *Science*, **235**, 458–460.

Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 7524–7528.

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

Chan, H. S. & Dill, K. A. (1990). The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **92**, 3118–3135.

Chan, H. S. & Dill, K. A. (1991). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 447–490.

Clore, G. M., Brünger, A. T., Karplus, M. & Gronenborn, A. M. (1986). Application of molecular dynamics with interproton distance restraints to 3D protein structure determination. *J. Mol. Biol.* **191**, 523–551.

Derrida, B. (1981). Random-energy model: an exactly solvable model of disordered systems. *Phys. Rev.* **24**, 2613–2624.

Elöve, G. A., Chaffotte, A. F., Roder, H. & Goldberg, M. E. (1992). Early steps in cytochrome *c* folding probed by time-resolved circular dichroism and fluorescence spectroscopy. *Biochemistry*, **31**, 6876–6883.

Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991). The energy landscapes and motions of proteins, *Science*, **254**, 1598–1603.

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

Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 4918–4922.

Gutin, A. M., Badretdinov, A. Y. & Finkelstein, A. V. (1992). Why are the statistics of globular protein structures Boltzmann-like? *Mol. Biol. (USSR)*, **26**, 94–102.

Head-Gordon, T. & Stillinger, F. H. (1993). Predicting polypeptide and protein structures from amino acid sequence: antlion method applied to melittin. *Biopolymers*, **33**, 293–303.

Hilhorst, H. J. & Deutch, J. M. (1975). Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume. *J. Chem. Phys.* **63**, 5153–5161.

Honeycutt, J. D. & Thirumalai, D. (1992). The nature of folded states of globular proteins. *Biopolymers*, **32**, 695–709.

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

Karplus, M. & Weaver, D. L. (1976). Protein-folding dynamics. *Nature (London)*, **260**, 404–406.

Karplus, M. & Weaver, D. L. (1979). Diffusion-collision model for protein folding. *Biopolymers*, **18**, 1421–1437.

Kim, P. & Baldwin, R. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**, 459–489.

Kim, P. & Baldwin, R. (1990). Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* **59**, 631–660.

Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 8721–8725.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44–45.

Levinthal, C. (1969). In *Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House, Monticello, IL* (Debrunner, P., Tsibris, J. C. M. & Münck, E., eds), pp. 22–24, University of Illinois Press, Urbana, IL.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

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

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

O'Toole, E. M. & Panagiotopoulos, A. Z. (1992). Monte Carlo simulation of folding transitions of simple model proteins using a chain growth algorithm. *J. Chem. Phys.* **97**, 8644–8652.

Piela, L., Kostrowicki, J. & Scheraga, H. A. (1989). The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *J. Phys. Chem.* **93**, 3339–3346.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.

Privalov, P. L. (1989). Thermodynamic problems of protein structure. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 47–69.

Privalov, P. L. & Khechinashvili, N. N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J. Mol. Biol.* **86**, 665–684.

Ptitsyn, O. B. (1987). Protein folding: hypotheses and experiments. *J. Protein Chem.* **6**, 273–293.

Ptitsyn, O. B. & Volkenstein, M. V. (1986). Protein structures and neutral theory of evolution. *J. Biomol. Struct. Dynam.* **4**, 137–156.

Roder, H. & Elöve, G. A. (1993). Early stages of protein folding. In *Issues in Protein Folding. Frontiers in Molecular Biology* (Pain, R. H., ed.), Academic Press, New York. In the press.

Rooman, M. J. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins. *Biochemistry*, **31**, 10239–10249.

Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry*, **31**, 10226–10238.

Šali, A. & Blundell, T. L. (1993). Comparative protein

modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.

Šali, A. (1991). Modelling three-dimensional structure of proteins from their sequence of amino acid residues. Ph.D. thesis, University of London, London.

Scheraga, H. A. (1989). Calculations of stable conformations of polypeptides, proteins, and protein complexes. *Chemica Scripta*, **29A**, 3–13.

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

Shakhnovich, E. I. & Gutin, A. M. (1989). Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys. Chem.* **34**, 187–199.

Shakhnovich, E. I. & Gutin, A. M. (1990*a*). Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature (London)*, **346**, 773–775.

Shakhnovich, E. I. & Gutin, A. M. (1990*b*). Enumeration of all compact conformations of copolymers with random sequence of links. *J. Chem. Phys.* **93**, 5967–5971.

Shakhnovich, E. I. & Gutin, A. M. (1993*a*). A new approach to the design of stable proteins. *Protein Eng.* In the press.

Shakhnovich, E. I. & Gutin, A. M. (1993*b*). Engineering of stable and fast-folding sequences of model proteins. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 7195–7199.

Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991). Protein folding bottlenecks: a lattice Monte Carlo simulation. *Phys. Rev. Letters*, **67**, 1665–1668.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*, **250**, 1121–1125.

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

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Nat. Acad. Sci., U.S.A.* **70**, 697–701.

Wetlaufer, D. B. & Ristow, S. (1973). Acquisition of three-dimensional structure of proteins. *Annu. Rev. Biochem.* **42**, 135–158.

Zwanzig, R., Szabo, A. & Bagchi, B. (1992). Levinthal's paradox. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 20–22.

Edited by B. Honig

(Received 25 March 1993; accepted 25 August 1993)