Reprinted from *J. Mol. Biol.* (1990) 211,

Dynamic Monte Carlo Simulations of Globular Protein Folding/Unfolding Pathways

II. α-Helical Motifs

Andrzej Sikorski and Jeffrey Skolnick

JMB285 JMB285A JMB285B JMB285T JMB285T2
CAM98
J. Mol. Biol. (1990) 211,

Dynamic Monte Carlo Simulations of Globular Protein Folding/Unfolding Pathways

II. α-Helical Motifs

Andrzej Sikorski† and Jeffrey Skolnick‡

Institute of Macromolecular Chemistry
Department of Chemistry, Washington University
St Louis, MO 63130, U.S.A.

(Received 4 April 1989, and in revised form 26 September 1989)

Dynamic Monte Carlo simulations of the folding pathways of α-helical protein motifs have been undertaken in the context of a diamond lattice model of globular proteins. The first question addressed is the nature of the assembly process of an α-helical hairpin. While the hairpin could, in principle, be formed *via* the diffusion-collision-adhesion of isolated performed helices, this is not the dominant mechanism of assembly found in the simulations. Rather, the helices that form native hairpins are constructed on-site, with folding initiating at or near the turn in almost all cases. Next, the folding/unfolding pathways of four-helix bundles having tight bends and one and two long loops in the native state are explored. Once again, an on-site construction mechanism of folding obtains, with a hairpin forming first, followed by the formation of a three-helix bundle, and finally the fourth helix of the native bundle assembles. Unfolding is essentially the reverse of folding. A simplified analytic theory is developed that reproduces the equilibrium folding transitions obtained from the simulations remarkably well and, for the dominant folding pathway, correctly identifies the intermediates seen in the simulations. The analytic theory provides the free energy along the reaction co-ordinate and identifies the transition state for all three motifs as being quite close to the native state, with three of the four helices assembled, and approximately one turn of the fourth helix in place. The transition state is separated from the native conformation by a free-energy barrier of mainly energetic origin and from the denatured state by a barrier of mainly entropic origin. The general features of the folding pathway seen in all variants of the model four-helix bundles are similar to those observed in the folding of β-barrel, Greek key proteins; this suggests that many of the qualitative aspects of folding are invariant to the particular native state topology and secondary structure.

1. Introduction

In spite of the extensive attention received during the last 15 years, the mechanism for the formation of the biologically active conformation of a globular protein is still far from understood (Kim & Baldwin, 1982; Anfinsen, 1972; Ghelis & Yon, 1982; Creighton, 1985, 1988). It is not easy to study the pathway of folding experimentally because the process of folding is highly co-operative and relatively fast (Creighton, 1988). Thus, studies of possible intermediate states in the folding process were relatively rare and mainly involved disulfide bridge trapping of intermediates (Creighton, 1974, 1977*a*,*b*, 1979, 1985, 1988). More recently, preliminary experimental studies have emerged employing, among other techniques, hydrogen exchange labeling methods (Roder *et al.*, 1988; Udgaonkar & Baldwin, 1988). However, the general mechanism of protein folding is not well established, nor are the general rules of protein folding well understood (Wright *et al.*, 1988). Recently, the nature of the protein folding process has been examined in a series of papers (Kolinski *et al.*, 1986*a*,*b*, 1987 Skolnick *et al.*, 1988, 1989*a*,*b*; Sikorski &Skolnick, 1989*a*,*b*) by means of Monte Carlo simulations. In the present work, these techniques are applied to study the mechanism of folding and unfolding of all the motifs

† Permanent address: Department of Chemistry, University of Warsaw, 02-093 Warsaw, Poland.
‡ Author to whom all correspondence should be addressed. Present address: Department of Molecular Biology, Research Institute of Scripps Clinic, 10666 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.

0022-2836/90/000000-00 $03.00/0

of four-helix bundles (Richardson, 1981; Ford *et al.*, 1984; Abdel-Meguid *et al.*, 1987); a similar analysis of the pathways of protein folding in Greek key, β-barrels appears in the accompanying paper, hereinafter designated Paper I (Skolnick & Kolinski, 1989).

The organization of this paper is as follows. Section 2 presents a brief description of the model. For more details, the reader is referred to section 2 of Paper I Skolnick & Kolinski, 1989). Section 3 begins with a discussion of the mechanism of assembly of an α-helical hairpin and addresses the question of whether or not a diffusion-collision-adhesion-model involving preformed helices describes the mechanism of assembly. Then, the nature of the folding/unfolding pathways of four-helix bundles having tight bends and one and two long loops is explored. The section concludes with a theoretical analysis of the free energy along the reaction co-ordinate of the four-helix bundle motifs having tight bends and one and two long loops. Section 4 summarizes the qualitative conclusions gleaned from these simulations.

2. Overview

(a) *Description of the Model*

The model chain consists of a consecutive sequence of n amino acid residues connected by $n-1$ bonds, confined to a tetrahedral (diamond) lattice. The model can be viewed as a crude, α-carbon representation of a protein chain. Because of lattice constraints, each α-helical turn consists of four residues (contrary to the 3·6 residues/turn in a real α-helix). A pair of adjacent α-helical stretches in the model helix bundle can only be antiparallel. This is to be contrasted with the case of real proteins, where the helices are not exactly parallel (the angle of crossing can be up to 20°) and may not be perfect; that is, they sometimes are broken into smaller pieces (Richardson, 1981; Janin & Chothia, 1980; Barlow & Thornton, 1988). Furthermore, on a diamond lattice, it is impossible to mimic the super-twist of the helix bundle that is observed in real proteins (Richardson, 1981; Weber & Salemme, 1980). Nevertheless, using this simple lattice model, one has a correct representation of the topologies of real proteins at low resolution, and thus these models can perhaps help elucidate the general rules about the pathways involved in the folding of tertiary structures.

In the model of helix-forming proteins, the helical wheel effect is implemented to accommodate short-range interactions as follows. If five consecutive residues form two g^- states (i.e. a single α-helical turn has been formed) and if this conformation is preferred, then the ith and $(i+4)$th beads interact with an attractive potential of mean force ε_c (a negative quantity). The addition of the next g^- state to this sequence leads to a decrease in free energy equal to ε_c.

We have defined the reduced temperature of the model system as $T^* = k_B T/\varepsilon_h$, where k_B is the Boltzmann's constant, T is the absolute temperature and ε_h is the hydrophilic residue pair interaction parameter. A further, more detailed description of the hydrophobic/hydrophilic pattern of interactions corresponding to amphipathic sequences is given by Sikorski & Skolnick (1989*a*).

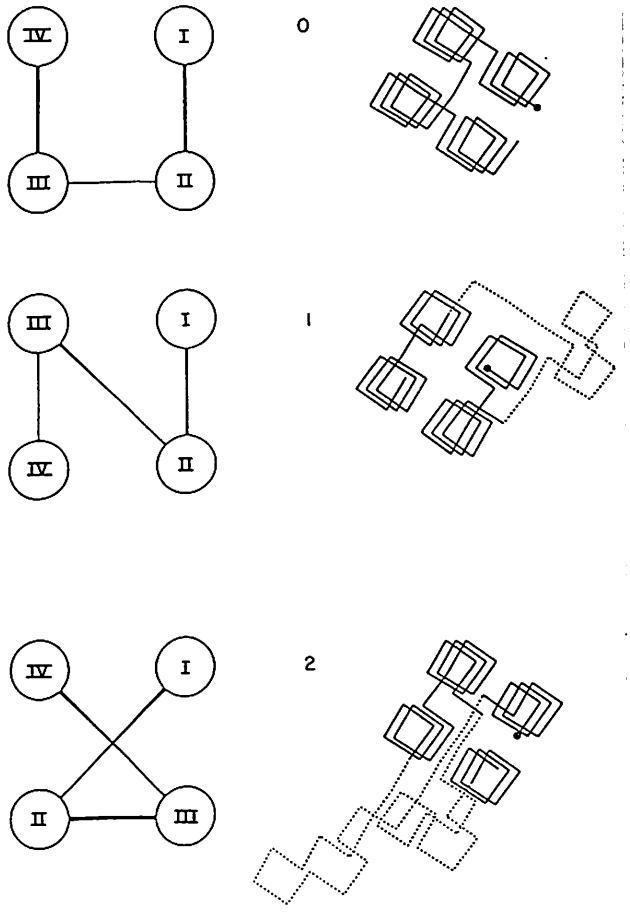


Figure 1. Representation of topologies (left) and the folded native configurations (right) for models 0, 1 and 2.

As in our previous studies, we employ the following shorthand notation to specify the primary sequence. $H_i(k)$ denotes a piece of the chain that can form the ith putative α-helical stretch in the native state and possesses the appropriate amphipathic sequence consisting of k residues. b_i^0 corresponds to a short bend neutral region containing three residues, and $L_i(k)$ indicates the ith putative loop region composed of k residues. We once again emphasize that this is only a primary sequence description, and there is no *a priori* restriction of the allowed conformations of the chain.

There are three variants of the four-helix bundle under consideration. *Model 0*, whose native state topology and folded conformation are shown on the right and left-hand sides of the top of Figure 1, consists of $n = 48$ amino acid residues. The native structure contains four almost identical α-helices (each containing almost 4 helical turns) connected

by three short bends (each bend involves 3 residues). The first turn (connecting helices I and II) consists of residues 11 through 13, the second turn (connecting helices II and III) consists of residues 23 through 27 and the third turn (connecting helices III and IV) consists of residues 35 through 37. Two different primary sequences were used for the case of model 0: (1) a central turn neutral region: $H_i(12)H_2(10)b_1^0H_3(11)H_4(12)$ (model On1) and (2) three neutral bends $H_i(10)b_1^0H_2(9)b_2^0H_3(9)b_3^0H_4(11)$ (model On3). When no b_1^0 is specified, the pattern in the putative bend region (in the native conformation) is assumed to be consistent with that in the adjacent helical strands H_i and H_{i+1}, (the 1st 2 residues belong to H_i and the last residue belongs to H_{i+1}).

In model 1, whose native state topology and folded conformation are shown on the left and right-hand sides, respectively, in the middle of Figure 1, the chain consists of $n = 61$ residues; the native conformation has virtually the same α-helical bundle as in model 0. The main difference between the two models lies in the topology of connections between the helices; see Figure 1. In model 1, helices II and III are connected by a long loop involving residues 23 through 38. The two remaining connections between the helices are short bends located at residues 10–11 and 46–48. Two distinct primary sequences have been studied for this motif: (1) model 1n having two-turn neutral, tight bend regions, $H_i(9)b_1^0H_2(10)L_1(16)H_3(9)b_2^0H_4(11)$; and (2) model 1h, whose primary sequence is entirely devoid of residues having a local preference for tight bend formation $H_1(11)H_2(11)L_1(16)H_3(11)H_4(12)$.

In model 2, whose native state topology and folded conformation are shown in the left and right-hand sides, respectively, at the bottom of Figure 1, the chain consists of $n = 78$ residues having two long loops and one short bend between helices II and III. The short bend involves residues 41 through 43, and the loops involve residues 13 through 30 and 54 through 67. The primary sequence studied for this model are: (1) model 2n, which possesses a single central turn neutral bend region, $H_1(12)L_1(18)H_2(9)b_1^0H_3(11)L_2(14)H_4(11)$; and (2) model 2h, which is devoid of a turn neutral region, $H_1(12)L_1(18)H_2(11)H_3(12)L_2(14)H_4(11)$.

In addition to the mean radius sequence gyration, $\langle S^2 \rangle$, defined in equation (3) of Skolnick & Kolinski (1989), another property of interest is the normalized helix content, θ_h, defined by:

$$\theta_h(T^*) = \frac{f(T^*) - f_{coil}}{f_{nat} - f_{coil}} \qquad (1)$$

where $f(T^*)$, f_{coil} and f_{nat} are the fraction of helical states at the temperature T^*, in the denatured state, and in the native configuration, respectively (Sikorski, & Skolnick, 1989*a*,b). θ_h in equation (1) assumes values close to zero in the denatured state and unity for the pure native state. It is assumed, for the sake of simplicity, that $f_{coil} = 1/3$. To make results of models 0, 1 and 2 comparable, values of θ_h are calculated only for the residues that form α-helical bundles. Values of f_{nat} are found to be 0·8464, 0·8667 and 0·8919 for models 0, 1 and 2, respectively.

3. Results

(a) *General considerations*

The main goal of these simulations is to elucidate the folding and unfolding pathways of four-helix bundles having different types of interhelix connectivities. Models having differences in primary structure (type "n" models versus type "h" models) have been studied to clarify the influence of short bends and long loops on the folding pathway, especially in the early stages of folding, and to find out if slight modifications of the primary sequence (which does not change the native structure) lead to different pathways. Simulations were made for the same set of interaction parameters as reported in our previous papers; namely, $\varepsilon_c = \varepsilon_a = \varepsilon_h/2 = -\varepsilon_w/2$ (Sikorski & Skolnick, 1989*a*,*b*). There, it was demonstrated that the folding transition has the same features over a broad range of parameter space. As shown below, similar behavior has been found for the folding pathway; changes in the ratio of interactions over a broad range of parameter space does not change the observed pathways. Thus, for the sake of brevity, only representative results are presented.

(b) *Folding of α-helical hairpins*

Because of lattice restrictions, pieces of secondary structures like β-strands and α-helices are immobile if only the local moves discussed in the previous section are used. Thus, the algorithm cannot move an assembled α-helix without dissolving it. (In practice, a 3-bond kink motion can effectively move β-strands about; see Paper I, section 2). In reality, of course, the solvent exerts Brownian forces (Chandrasekhar, 1943), which can move assembled secondary structural elements; it is precisely this mechanism that the diffusion-collision-adhesion model involving helix preformation invokes as the dominent feature of tertiary structure assembly (Karplus & Weaver, 1976, 1979; Lee *et al.*, 1987). Thus, in order to examine the validity of the diffusion-collision-adhesion model, we have used a modified version of the algorithm that explicitly allows rigid helix rotations and translations.

The Monte Carlo simulation algorithm that has been applied to α-helical hairpins contains in addition to the three-bond, four-bond and end flips, two additional kinds of motions. One of them, shown in Figure 2(a), translates an α-helix in a random direction a distance of $8^{1/2}$; (this is the smallest possible distance an α–helix can be translated on the diamond lattice). The translation also slightly distorts one end of the helix but does not affect the other end. The shorter of the two random coiled tails attached to the helix is translated as well. The

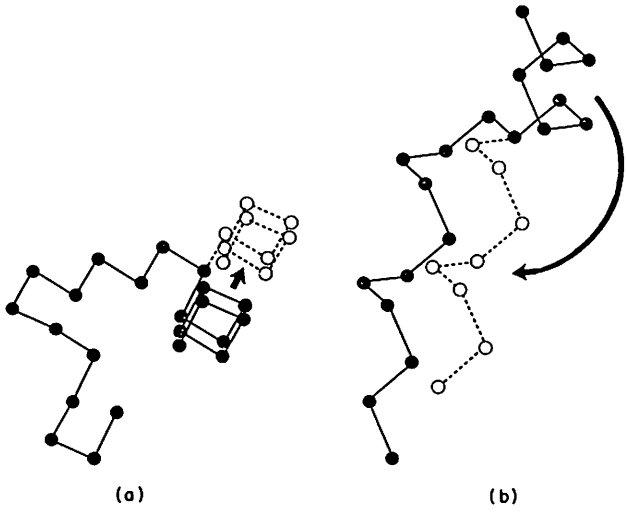


Figure 2. Example of long-distance moves of secondary structures involving the translation of (a) a helix and (b) rotation of a helix.

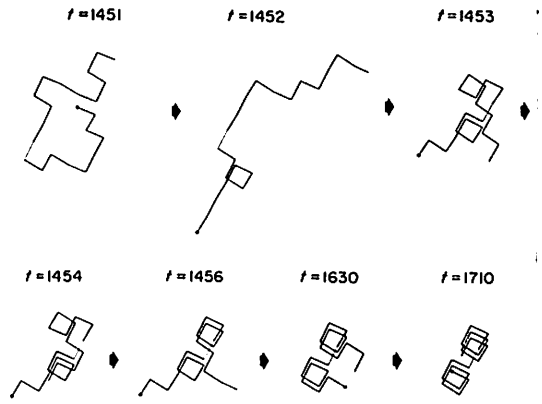


Figure 3. A representative folding trajectory of a single hairpin using the extended algorithm with the possibility of translations and rotations of helical fragments (the hydrophobic/hydrophilic pattern is the same as in the 1-2 hairpin of model 0n3 with a neutral bend) at a temperature $T^* = 0{\cdot}5$.

second move is a rotation of a helix. As shown in Figure 3(b), the entire helix (and shorter tail) is rotated around the first residue located outside the helix. Due to lattice restrictions, the angle of rotation can take only the values $2\pi/3$ or $4\pi/3$.

In order to include these new motions without significantly distorting the time-scale (they involve the motion of a larger number of beads than a 3 or 4-bond motion), these moves have to be penalized relative to the elemental three and four-bond jumps by weighing them appropriately. The main idea is that any translation or rotation can be treated as a combination of three-bond kink motions (which form the basis of the time unit). In a single three-bond kink motion (see Fig. 3(a) of Paper I), two residues are shifted by a distance of $8^{1/2}$. The probability of a single three-bond motion in a particular direction is equal to 0·5, because there are two different possibilities for this move. Consider a helix of r residues. Thus, the simplest approximation for the frequency of the attempted translation of a helix is given by $(1/2)^{r/2}$ (there are $r/2$ 3-bond motions in this same direction, because each 3-bond motion shifts 2 residues). Similarly, the probability of a rotation can be decomposed into the probability of a consecutive sequence of the three-bond motions. Here, however, the distance that each residue in the rotated helix moves is different. The frequency of rotation of a helix can be written as $(1/2)q$, where

$$q = 0{\cdot}5 \sum_{i=1}^{r} k_i,$$

where k_i is the number of three-bond motions required to rotate residue i. By employing the above, rigid body helix rotations and translations are implemented into the algorithm by penalizing them at every time step by the appropriate factor. This new algorithm, which in principle permits diffusion-collision of preformed helices to occur, has been applied to the simplest model system, a hairpin with a short bend between helices and that is composed of 22 residues having the primary sequence $H(9)b_1^0\mathrm{OH}(10)$. In Figure 3, we present a representative folding trajectory of the α-helical hairpin from the denatured state. As is clearly seen, the hairpin does not form by the formation of isolated helices which then diffuse together as the diffusion-collision-adhesion model conjectures. Rather, folding initiates by the formation of one or two hydrophobic contacts near the native turn as conjectured by Matheson & Scheraga (1978), and then the helices are constructed on-site. These rotations and translations of helical fragments do occur, but these seem to be rather rare events. Because they are marginally stable prior to assembly of the native state (the denatured state has about 30% helix content), the translated and rotated, isolated helices are not long-lived and undergo dissolution by micromodification. In other words, it is faster and more probable to assemble them on-site (which is downhill in free energy after the 1st few steps), than to wait for the isolated helices to form and then diffuse together. This process is too slow relative to the mean lifetime of the isolated helices.

To avoid underestimating the role of relative helix diffusion, we have artificially assigned the probability factors described above to have higher (and non-physical) values, up to the square-root of their original values. It should be pointed out that these probabilities are already overestimated by neglecting the constraints caused by one (shorter) tail that is translated (rotated) with the helix without incurring any additional penalty. This dramatic increase in the probability of moving the α-helices does not change the mechanism of folding, and the hairpin is again assembled step-by-step on-

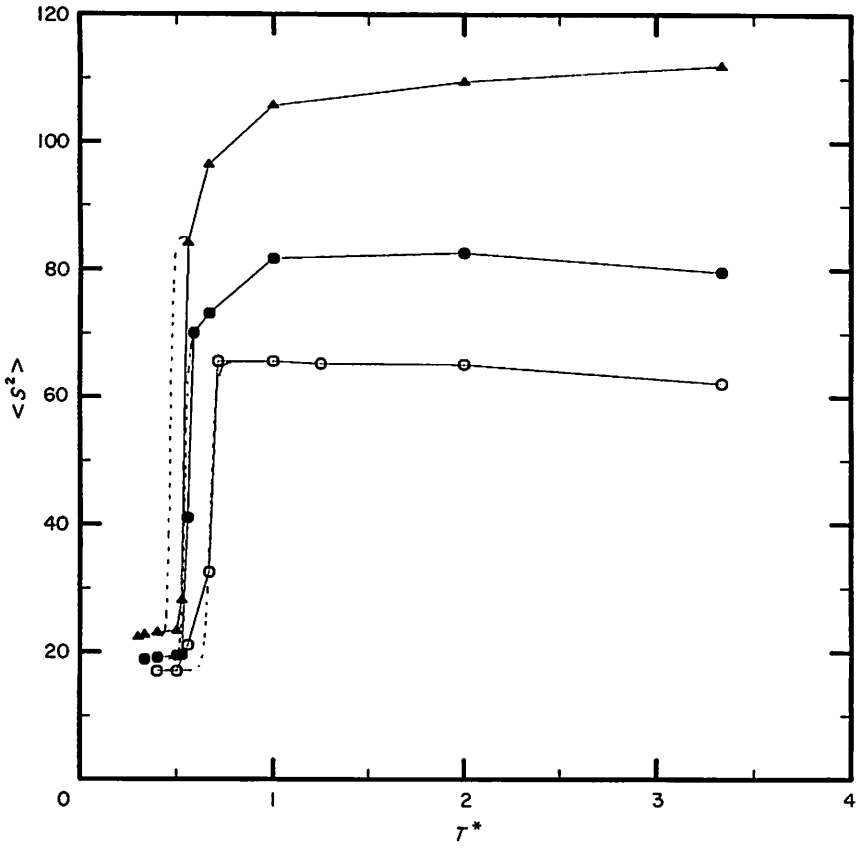


Figure 4. Plot of the mean-square radius of gyration $\langle S^2 \rangle$ *versus* the reduced temperature T^* averaged over folding/unfolding sequences. (Model 0n1 (open circles), model 1n (filled circles) and model 2n (open triangles). The broken curves correspond to results obtained from the analytical model, according to eqns (5) and (6).

site as in Figure 3. If the rotations and translations are not penalized at all relative to the three and four-bond kind motions, but are equally weighted, then assembly predominantly occurs by diffusion-collision-adhesion of preformed helices. In this non-physical case, there is a competition between these two mechanisms of folding, but for realistic estimates of the probability of rigid translations and rotations, the on-site construction mechanism is the dominant one. (This does, however, suggest that rigid rotations and translations be incorporated into equilibrium sampling algorithms where the pathways need not be physical.) Therefore, in what follows, we return to the original algorithm possessing only local three and four-bond motions and chain end modifications, as these are responsible for the assembly of native structure in these models.

(c) *Folding of four-helix bundles*

The analysis of the equilibrium properties of all models has been fully discussed in our previous work (Sikorski & Skolnick, 1989*a*,*b*). Here, we present only representative equilibrium results as a function of temperature. Figure 4 presents results for the temperature dependence of the mean-square radius of gyration, $\langle S^2 \rangle$, for models 0n1 (open circles), 1n (filled circles) and 2n (open triangles). $\langle S^2 \rangle$ changes its value on refolding because the native conformation of the protein is compact and densely packed, unlike the random coil state. In the case of the presence of long loop(s), $\langle S^2 \rangle$ weakly decreases even at temperatures below the transition regime, because only the native helical bundle is stable; the loop(s) undergo further small-scale rearrangements on cooling.

The folding statistics are compiled in Tables 1, 2 and 3 for models 0, 1 and 2, respectively. For every single simulation run, the total number of Monte Carlo steps is given in column two. Column three gives T^*, the temperature of the refolding (unfolding) transition for which the pathway of folding (refolding) pathway is analyzed. The conformational properties, averaged over these steps, are given in column four for the mean energy per residue in units of $k_B T$. Column five presents the mean-square radius of gyration $\langle S^2 \rangle$, and column six gives the helix content, calculated according to equation (1). Column seven describes the changes in conformation, and column eight gives a short description of the pathway of folding (refolding) by listing the number of helices in the sequence in which they assembled. Column nine presents the fraction of total simulation time for which a native hairpin was stable independent of whether folding is

Table 1
Summary of conformational properties of model 0

Run	$t/10^5$	T^*	$\langle E/n \rangle$	$\langle S^2 \rangle$	$\langle \theta_h \rangle$	Transition	Pathway	τ_5/t	τ_{10}/t	τ_5^*†	τ_{10}^*‡	τ_N^*§
A. *Model 03n*												
1	2	0·667	−0·2151	67·95	0·242	D	—	0·075	0	—	—	—
2	2	0·667	−0·8564	31·22	0·769	D→N	(3→4)→2→1	0·040	0·015			
3	2	0·667	−0·7092	36·55	0·690	D→N	(3→4)→2→1	0·090	0·0130	1000	400	800
4	2	0·625	−0·8976	26·44	0·866	D→N	(3→4)→2→1	0·030	0·030	1000	600	6200
5	2	0·625	−1·0069	23·36	0·815	D→N	(1→2)→3→4	0·015	0·030	800	1400	5000
6	2	0·667	−0·2166	66·08	0·263	D	—	0·105	0·030	—	—	—
7	2	0·625	0·7976	41·36	0·684	D→N	(2→3)→1→4	0·075	0·010	300	2000	2100
8	2	0·625	−1·1038	18·54	0·973	D→N	(4→3)→2→1	0·015	0·060	—	—	—
9	2	0·625	−0·9938	24·16	0·895	D→N	(3→4)→4→1	0·025	0·050	1000	2200	9400
10	2	0·625	−0·5310	45·06	0·542	D→N	(2→3)→1→4	0·135	0·030	800	1400	9000
11	2·5	0·769	−0·2500	59·57	0·283	N→D→N	(1→4)→2→3					
12	2·5	0·769	−0·9701	17·07	0·966	N	—	—	—	—	—	—
13	2·5	0·769	−0·3732	51·72	0·375	N→D	(1→4)→3→2					
14	2·5	0·769	−0·9686	17·08	0·966	N	—	—	—	—	—	—
15	2·0	0·667	−0·6653	41·56	0·6201	D→N→D	(1→2)→3→4	0·060	0·020	900	1800	700
							4→1→(2→3)			200	500	600
B. *Model 01n*												
1	2·5	0·667	−0·6661	39·84	0·621	D→N	(2→3)→1→4	0·012	0·020	500	500	1000
2	2	0·667	−0·8630	30·27	0·810	D→N	(3→4)→2→1	0·060	0			
3	2	0·667	−0·2539	67·57	0·302	D	—	0·060	0·010	—	—	—
4	2	0·667	−0·8682	28·30	0·833	D→N	(2→3)→4→1	0·050	0·050	1000	1250	9000
5	2	0·667	−0·9571	27·32	0·857	D→N	(2→3)→1→4	0·040	0·020	400	400	3000
6	2	0·667	−0·2584	66·29	0·334	D	—	0·072	0·024	—	—	—
7	2	0·667	−0·5934	45·57	0·579	D→N	(1→2)→3→4	0·085	0·035			
8	2	0·667	−0·4289	55·36	0·470	D→N	(2→3)→4→1	0·040	0·025	100	3500	2300
9	2	0·667	−1·0907	18·77	0·973	D→N→D	(2→3)→4→1	0·010	0·085	750	3250	17,250
10	2	0·667	−1·0176	24·34	0·901	D→N	(2→3)→1→4	0·010	0·005			
11	2	0·667	−0·7446	37·87	0·713	D→N	(2→3)→4→1	0·030	0·045	250	2500	7750
12	2·5	0·769	−0·5924	36·18	0·655	N→D	4→1→(2→3)	0·020	0·001			
13	2·5	0·833	−1·1414	61·15	0·368	N→D	4→1→(2→3)	0·010	0·020	1450	750	300
14	2·5	0·769	−0.8637	24·29	0·852	N→D		0·020	0·080			
15	1	0·769	−0·8463	31·07	0·787	D→N	(2→3)→1→4	0·120	0·012	750	1250	7250

† For a folding transition, τ_5^* is the time from initiation of folding to the formation of a helical hairpin, given that a successful folding event will occur. For unfolding, τ_5^* is the time from the 1st appearance of the hairpin until the last remaining native contacts have dissolved, given that unfolding has occurred from the N state.

‡ For a successful folding transition τ_{10}^* is the time from the 1st appearance of the hairpin until the 3-helix bundle has formed. For an unfolding transition, τ_{10}^* is the time from the 1st appearance of the 3-member barrel until the 1st appearance of the helical hairpin, i.e. it is the time for dissolution of 1 of the helices in the 3-member bundle, given that unfolding of the entire molecule will occur.

§ For a successful folding transition τ_N^* is the time from the 1st appearance of the 3-helix bundle until the native state is formed. For an unfolding transition, τ_N^* is the time from the initiation of unfolding until the 3-helix bundle 1st appears, given that 1 of the 4 helices will completely dissolve, and that unfolding of the entire molecule will occur.

successful. Column ten presents the ratio of the lifetime of the three-member bundle intermediate state to the total simulation time, independent of whether folding is successful. Column 11 presents the time from the initiation of folding to the first appearance of a native hairpin, τ_5^*, given that a successful folding event will occur. Column 12 gives the elapsed time for assembly of a three-helix bundle (the 2nd intermediate state), from the formation of the hairpin, τ_{10}. Column 13 gives the time of assembly of the native state from three-member bundle, τ_N.

Figure 5, presents a plot of the instantaneous value of the mean-square radius of gyration, S^2, *versus* time for model 0n1 at $T^* = 0{\cdot}667$, corresponding to the transition region extracted from the 11th simulation run). For the first 3×10^4 time steps, the chain has not collapsed, and its dimensions, characteristic of the denatured state, strongly fluctuate between the values 30 and 120. After 3×10^4 steps, the radius of gyration diminishes and rapidly approaches values close to 20, and the amplitude of the fluctuations become much lower. This suggests that an ordered structure has formed, but the dimensions and size of the fluctuations indicate that it is not yet in the native state. When $t = 3{\cdot}8 \times 10^4$, the radius of gyration becomes lower ($S^2 = 17{\cdot}0$), assuming native-like values with small-scale fluctuations. Figure 6 shows the change in the number of instantaneous native contact pairs, N_C, *versus* time for the same simulation run. The maximum value of N_C, in the native configuration, is 21, for the model under consideration. One can see that during the first portion of the simulation for times up to 3×10^4, the number of native contacts fluctuates between 0 and 5. Five native contacts indicate that the formation of a hairpin has occurred. After 3×10^4 time units, the number of native

Table 2

Summary of conformational properties of model 1

Run	$t/10^5$	T^*	$\langle E/n \rangle$	$\langle S^2 \rangle$	$\langle Q_h \rangle$	Transition	Pathway	τ_5/t	τ_{10}/t	τ_5^*†	τ_{10}^*‡	τ_{NS}^*§
A. *Model 1n*												
1	2·5	0·556	−0·4075	56·71	0·698	D	—	0·440	0·100	—	—	—
2	2·5	0·526	−0·5187	55·50	0·769	D	—	0·420	0·300	—	—	—
3	2·5	0·526	−0·8280	30·04	0·921	D→N	(1→2)→3→4	0·140	0·140			
4	2	0·526	−0·9681	20·66	0·923	D→N	(3→4)→2→1	0	0·704			
5	6·5	0·526	−0·7193	52·83	0·907	D→N	(1→2)→3→4	0·410	0·225	500	3750	13,000
6	1·0	0·526	−1·1050	22·90	0·979	D→N	(1→2)→3→4	0·020	0·060	250	18,500	5250
7	5	0·526	−0·6827	46·16	0·818	D→N	(3→4)→2→1	0·320	0·220	1000	3750	10,750
8	5	0·526	−0·7071	45·09	0·816	D→N	(3→4)→2→1	0·280	0·140	3100	13,600	24,000
9	2·5	0·588	−1·0282	20·62	0·983	N	—	—	—	—	—	—
10	1	0·588	−0·7079	35·92	0·768	N→D	(1→2)→(3→4)	—	—	100	1200	2800
11	1·0	0·528	−1·1535	19·58	0·987	N	—	—	—	—	—	—
B. *Model 1h*												
1	5	0·526	−0·4628	63·62	0·751	D	—	0·400	0·120	—	—	—
2	6	0·538	−0·4124	67·96	0·731	D	—	0·380	0·040	—	—	—
3	2·5	0·526	−0·8048	42·12	0·893	D→N	(2→4)→1→3	0·160	0·120	450	400	18,375
4	4	0·526	−0·7503	39·73	0·889	D→N	(2→4)→1→3	0·120	0·300	750	4500	39,750
5	4	0·588	−1·0101	19·92	0·979	N	—	—	—	—	—	—
6	4	0·588	−1·0163	19·97	0·986	N	—	—	—	—	—	—
7	2·5	0·641	−0·3742	62·91	0·674	N→D	(4→1)→2→3	0·020	0·020	250	2250	200
8	3	0·500	−1·0150	28·29	0·925	D→N	(1→2)→3→4	0·160	0·080	500	6000	25,000
9	5	0·526	−0·6520	44·88	0·841	D→N	(1→2)→3→4	0·010	0·220	800	3800	102,400
10	8	0·556	−0·9602	48·96	0·818	D→N	(3→4)→2→1	0·180	0·030	250	5500	2250

For details, see the footnotes to Table 1.

contact pairs increases rapidly to 10 and then fluctuates above this value. This corresponds to the formation of a three-helix bundle, and fluctuates about this value. This corresponds to the formation of a three-helix bundle, and the location of this structure in time is the same as that of ordered structure in Figure 5. Hence, for a time-interval of about 3×10^4 steps, a transition from the denatured state, D, to the intermediate state, I, has occurred. The intermediate state is short-lived, and the

Table 3

Summary of conformational properties of model 2

Run	$t/10^5$	T^*	$\langle E/n \rangle$	$\langle S^2 \rangle$	$\langle Q_h \rangle$	Transition	Pathway	τ_5/t	τ_{10}/t	τ_5^*†	τ_{10}^*‡	τ_{NS}^*§
A. *Model 2n*												
1	5	0·526	−0·3345	76·84	0·349	D	—	0·220	0·010	—	—	—
2	5	0·526	−0·3638	71·71	0·472	D	—	0·240	0·020	—	—	—
3	5	0·521	−1·0362	25·29	0·977	D→N	(2→3)→1→4	0·000	0·005	1000	2000	23,000
4	5	0·526	−0·8770	29·65	0·932	D→N	(2→3)→1→4	0·110	0·040			
5	10	0·526	−0·9150	22·99	0·979	D→N	—	—	—			
6	10	0·526	−0·6287	46·16	0·818	D→N	(2→3)→4→1	0·35	0·210			
7	5	0·526	−0·7072	45·09	0·816	D→N						
8	5	0·588	−0·8430	23·65	0·842	N	—	—	—	—	—	—
9	5	0·588	−0·3709	79·79	0·585	N→D	(1→4)→(2→3)	—	—	1400	2400	4000
10	5	0·588	−0·8548	23·42	0·859	N	—	—	—	—	—	—
11	1	0·625	−0·7622	28·95	0·851	N→D	1→4→(2→3)	0·180	0·080	750	22,750	22,750
12	7·5	0·526	−0·5799	48·80	0·637	D→N	(2→3)→4→1	0·134	0·231	1000	2500	5500
B. *Model 2h*												
1	5	0·521	−0·5132	57·01	0·663	D→N	(2→3)→4→1	0·240	0·008	2000	6000	32,000
2	5	0·500	−0·9822	24·89	0·963	D→N	(2→3)→4→1	0·020	0·008			
3	1	0·625	−0·3798	27·75	0·8499	N→D	1→4→(2→3)	0·060	0·003	1000	2500	5000
4	5	0·526	−0·3038	87·32	0·492	D		0·400	0·000	—	—	—
5	1	0·588	−0·2581	80·14	0·214	N→D	(4→1)→(2→3)			200	200	2400
6	1	0·625	−0·7786	23·58	0·897	N	—	—	—	—	—	—
7	1	0·588	−0·6412	42·74	0·648	N→D	4→1→(2→3)	0·200	0·080			
8	1	0·625	−0·7966	23·82	0·948	N	—	—	—	—	—	—
9	8	0·521	−0·5392	51·57	0·658	D→N	(2→3)→1→4	0·025	0·073			
10	4	0·521	−0·8544	33·09	0·902	D→N	(2→3)→4→1	0·080	0·080			

For details, see the footnotes to Table 1.

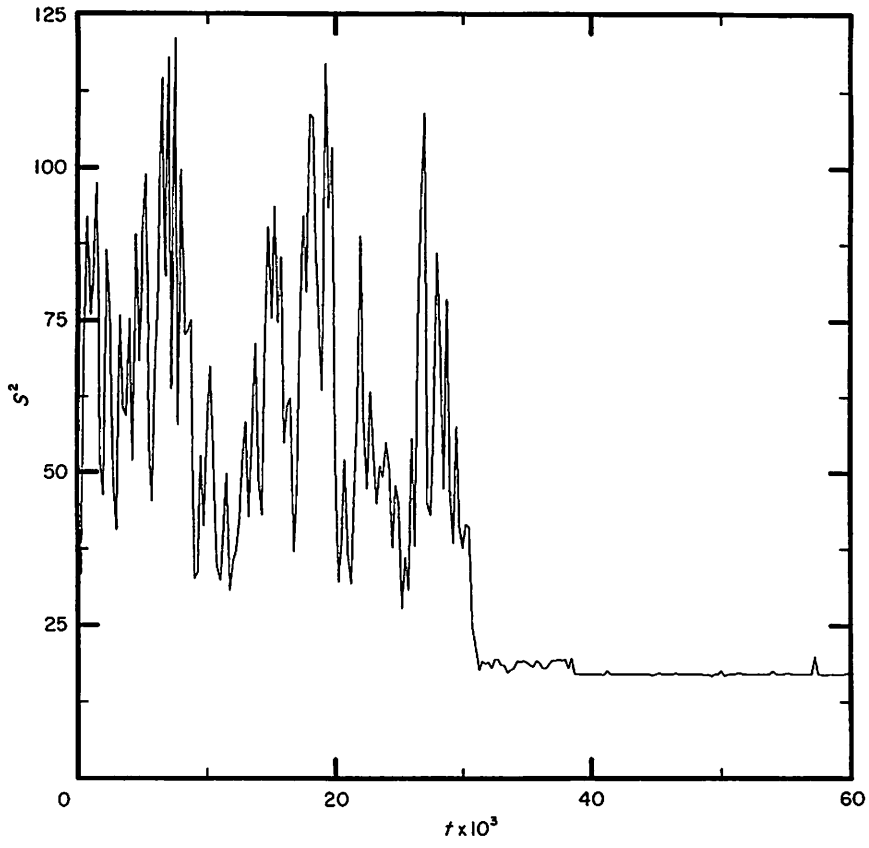


Figure 5. Plot of the instantaneous radius of gyration *versus* time for model 0n1 at a temperature $T^* = 0{\cdot}667$.

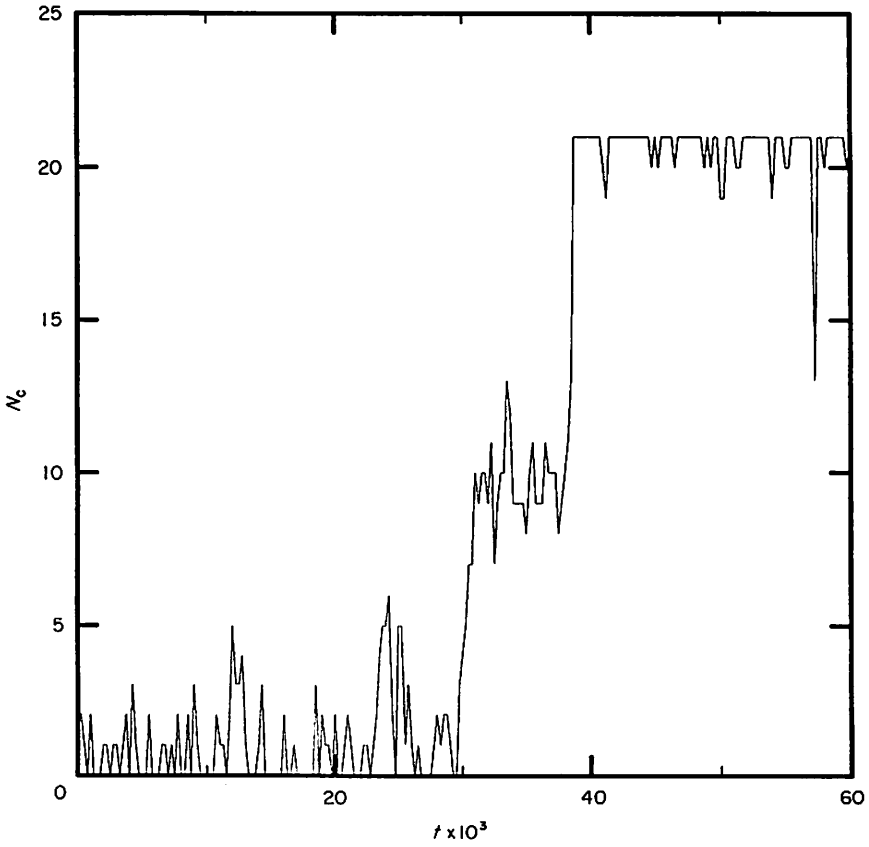


Figure 6. Plot of the instantaneous number of native contacts *versus* time for the model 0n1 at a temperature $T^* = 0{\cdot}667$.

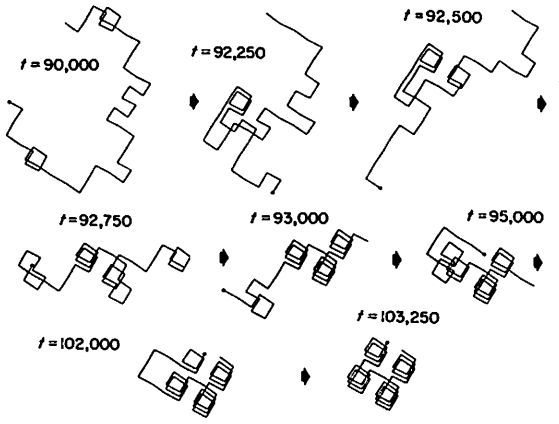


Figure 7. Representative folding trajectory for model 0n1 (at a temperature ($T^* = 0{\cdot}667$).

second transition to the native state, N, occurs; this is indicated by the very rapid increase of the number of native contact pairs up to 21, at a time $= 3{\cdot}8 \times 10^4$. After this time, the number of contacts fluctuates, basically between 19 and 21. This suggests that only the ends of the chain can change conformation, and the native state is stable for the conditions under consideration. It is impossible to find out precisely from Figures 5 and 6 what the pathway and mechanism of folding are. This question can be answered by a detailed analysis of the folding trajectories.

In Figures 5 and 6, a single D → I → N transition has been shown and discussed. For temperatures where N and D are equilibrium, the model chain undergoes a series of folding and unfolding transitions at any moment in time, the entire chain is almost always randomly coiled or in the native state, and the time spent in intermediates is small compared to that of a native state. This means that the folding transition has an all-or-none character, as in real proteins (Tanford, 1968; Privalov, 1979). The population of the native state in the transition region increases with decreasing temperature, and eventually, only the native state is stable at lower temperatures.

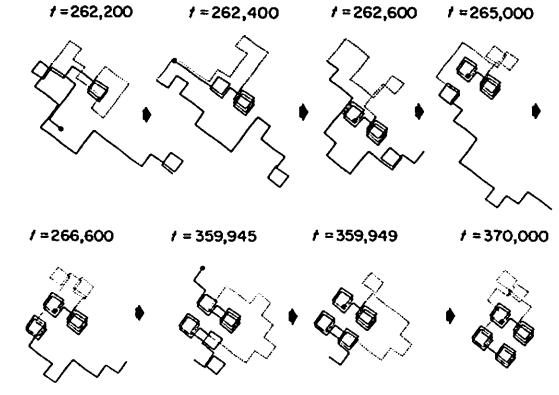


Figure 8. A representative folding trajectory for model 1h (at a temperature $T^* = 0{\cdot}556$).

It should be pointed out that the behavior described above is common to all three topologies under consideration. There is no qualitative difference between the behavior of S^2 and N_C *versus* time for models 0, 1 and 2. Of course, the introduction of a long loop(s) makes the folding time relatively longer, and intermediate species (hairpins and 3-helix bundles) are more populated. This is because the formation of a three-helix bundle intermediate involves a huge reduction of loop(s) entropy (Jacobson & Stockmayer, 1950; Flory, 1956). The population of hairpins is larger for models 1 and 2, because their folding transition occurs at lower temperatures, where a hairpin is more stable. The presence (or lack) of neutral bends does not make a significant difference in the case of models with long loops (1n, 1h, 2n and 2h), and the temperatures of folding differ only slightly (this effect is discussed by Sikorski & Skolnick, 1989*b*). There is no essential difference between the conformational transition of models On1 and On3; for both models, the unique native structure (Fig. 1(a)) is obtained. The discussion of the equilibrium effect of short bends and long loops is presented elsewhere (Sikorski & Skolnick, 1989*b*)

It is, of course, possible to monitor the configurations of the chain during the simulation runs, and thus check what the folding pathways are explicitly. This has to be repeated many times to determine whether well-defined folding pathways exist. Due to limitations in computer time, we were able to examine about 20 folding events for each topology. In the series of pictures presented in Figures 7, 8 and 9, representative folding trajectories for models 0, 1 and 2 are shown. Generally, the number of different configurations during a single Monte Carlo run is very large (of the order of 10^5 to 10^6), but

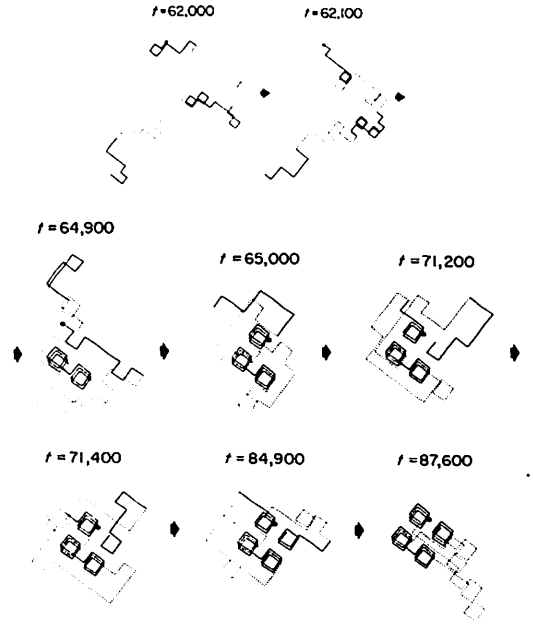


Figure 9. A representative folding trajectory for model 2h (at a temperature $T^* = 0{\cdot}528$).

only a few configurations are interesting because they are important for the folding pathway; thus we have selected the relevant folding events. During the time between the displayed snapshots, the model chain remains in very similar configurations. For example in Figure 9, between times $t = 266{,}600$ and $t = 359{,}945$, the motif formed from α-helical strands 1, 2 and 3 is preserved, and the remaining part of the chain hunts through configuration space without forming any additional, stable, long-lived secondary structure(s).

It is immediately apparent from Figures 7, 8 and 9 that all three models, having different native topologies, undergo folding along very similar pathways. There are a number of distinguishable stages in the process, common to all models under consideration. The first order motif that forms from a denaturated protein chain is a hairpin or eyelet. We use the term hairpin to describe two antiparallel α-helical stretches connected by a short bend, and the term eyelet to describe a pair of antiparallel helices joined together by a long structureless loop. The latter motif is observed with considerably lower frequency in the case of models 0 and 1. Hairpins and eyelets are relatively stable (the hairpin is of course more stable) but, during longer periods of time, they can dissolve as well. In the case of model On1, the central hairpin (formed from helices 2 and 3) is predominantly formed (see the configurations between $t = 90{,}000$ and $t = 92{,}750$ in Fig. 7). Introduction of three neutral bends (as in model 0n3) leads to the more frequent occurrence of the two remaining hairpins (1-2 and 3-4). Hence the hairpins consisting of the pair of helical stretches 1-2, 2-3 or 3-4 can lead to the native state for model 0 (Table 1, column 8). In the case of models 1 and 2, the presence of tight, neutral bends has no influence on hairpin formation (Sikorski & Skolnick, 1989*b*). In model 1, hairpins 1-2 or 3-4 lead to native states, and in model 2, only the 2-3 hairpin lead to the native state. A helical hairpin in models 1 and 2 is formed in the same way as described above; see Figure 8 ($t < 262{,}600$) and Figure 9 ($t < 64{,}900$). For models having long loops, it is possible to get the native state from an eyelet of the 1-3 or 2-4 type, but the probability of this pathway is relatively small.

The formation of α-helical hairpin (eyelet) can start from any residue (see below), but the most frequently observed initiation of folding occurred in the vicinity of a bend. At the initiation of folding, one or two hydrophobic contacts are formed, and then, or sometimes simultaneously, the formation of one or two helical turns occurs. This resembles the hydrophobic pocket formation in the early stages of folding that was conjectured by Matheson & Scheraga (1978). The formation of a few hydrophobic contacts probably partially compensates for the loss in configurational entropy (Flory, 1956; Jacobson & Stockmayer, 1950). Very quickly following the formation of the first few contacts, the helical stretches of the remaining hairpin form. At other times, the process of hairpin assembly can be more asymmetric; that is, almost an entire helical stretch and only one (usually close to the bend) helical turn of the second stretch is formed at first. Thus, it appears that the formation of α-helical hairpins is a co-operative event in which both short and long-range interactions are involved.

It should be pointed out, however, that there is a difference in the stabilization energy of α-helical and β-strand hairpins. From Table 5 in Paper I (Skolnick & Kolinski, 1989), we find that the energy of the 2-3 hairpin is $-12{\cdot}3\ k_BT$ at the transition temperature ($T^* = 1{\cdot}237$). The energy of the first folded helical hairpin at the transition temperature is $-17{\cdot}8\ k_BT$ for model 0, $-23{\cdot}1\ k_BT$ for model 1 and $-24{\cdot}0\ k_BT$ for model 2. The loss of the configurational entropy associated with hairpin formation at a given strand length is virtually the same for both α-helical and β-strand cases (see section 3(e)). In the case of an α-helical hairpin, the much larger decrease in energy arising from hydrophobic contacts can override the entropy barrier. Thus, the formation of a α-helical hairpin starts from a contact pair down the stretch as well as from a bend. When combined with the hairpin assembly results of (section 3(b)), these simulations provide no evidence for microhelical domain coalescence as the mechanism of hairpin assembly. Because the simulations are performed at low temperature, there is a relatively large amount of α-helical states in denatured state, but these helices are not involved in the forming of tertiary structures; that is, the helices that lead to successful tertiary structure formation are constructed on-site. There is only one center of initiation, and both helices are formed from this single centre.

Folding from a random coli does not have to lead directly to an in-register hairpin. It is possible that out-of-register structures are assembled. There are distorted hairpins with non-native (hydrophobic) contacts, helices whose length is greater than that of a native stretch, which are stabilized by hydrophobic contacts (especially when there is no intrinsic bend preference between 2 native helical stretches) and mirror image hairpins. All these structures are located in local minima on the free-energy surface, and a model chain has to spend some time trapped there. If the system is considerably quenched, it is possible that the model chain can be trapped for times at least comparable to that of the simulation and a native state will not be obtained.

The next step along the folding pathway is the formation of a three-member, α-helical bundle that involves the addition of the third helix to an existing hairpin. This is the relatively long-lived folding intermediate (see Tables 1, 2 and 3, columns 11-13). In the case of model 0, this stage of folding begins with the formation of the native bend at one of the ends of hairpin, and successively the helical turns are formed; the next helix zips up. For models 1 and 2, the situation is different because the third helix to be formed is attached to the hairpin by a long loop. Because the entropic barrier to be

surmounted is higher, the life-time of the helical hairpin is longer. There is no substantial difference between the entropy of assembly starting from any point in the third α-helical stretch; hence, one can observe that the process of the formation of the third helix can start from anywhere in this helix, with possible subsequent rearrangement to insure registration. There are many equally probable states involving the loop plus nascent and perhaps not-in-register helix; this effect is mainly responsible for the pause following formation of the hairpin prior to correct third helix assembly. Out-of-register helices can be eliminated by bubble formation that produces registration and/or complete dissolution of the non-native helix.

The final step in assembly corresponds to the addition of the last helix to the bundle. For model 0, this step resembles the previous one. The fourth helix starts to fold beginning from the bend and zips up (e.g. Fig. 7, $t = 93{,}000$). In the case of models 1 and 2, the formation of the last helix must also surmount a higher entropic barrier; the chain that forms the fourth helix must snake its way up to a channel formed by the first propagation and stabilizes the three-member, helical bundle but, on the other hand, long, coiled loop(s) located close to the bundle restrict the ways of assembling the native state due to the loop's excluded volume. Hence, the life-time of the intermediate species τ_N^s (3-member bundle) is longer here than for model 0; nevertheless, this state is still much less populated compared to the native and denaturated states. This means that the process of folding retains its all-or-none character for protein models containing long loops.

There is no single folding pathway for the models under consideration. For all topologies, a large diversity of folding pathways is possible in principle but, due to the configurational entropic barriers, some of them virtually do not occur (e.g. assembling *via* eyelets). In the beginning, the number of possible pathways is the largest and decreases the closer one is to the native state. For example, if we consider model 0, one of three hairpins can be formed, then there are two ways of assembling the third helix, and finally there is only one way to form the native state from the three-member bundle.

(d) *Unfolding pathways of four-helix bundles*

In Figure 10, a representative unfolding pathway is presented for model 2h. One can see that, in general, it is the reverse of the folding pathway; one of the end helical stretches (1 or 4) dissolves, starting from a chain end or from the side of long loop, as is shown for $t = 87{,}500$ and 88.000. The remaining folded structure, the three-helix bundle, is relatively stable and forms a long-lived intermediate. Then the next helix (that connected by a long loop) starts to dissolve from the end, in a similar way as the first one, leaving only a single hairpin, the hairpin lives for relatively less time than the three-member helical bundle, and

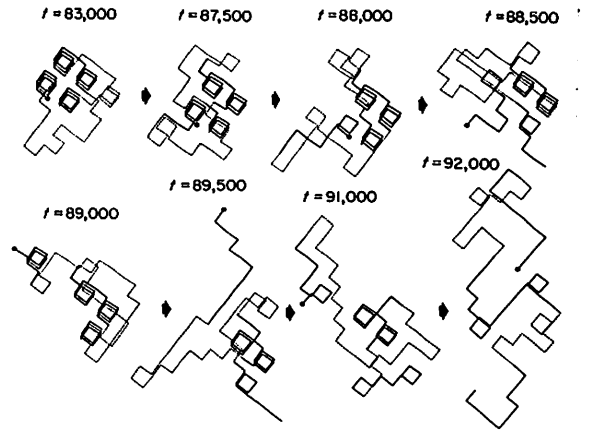


Figure 10. A representative unfolding trajectory for model 2h (at a temperature $T^* = 0{\cdot}556$).

eventually dissolves. Similar to the folding process, there is more than one pathway of unfolding. In the case of unfolding transitions at slightly higher temperatures, the unfolding process is very fast, and helices 1 and 4 dissolve virtually simultaneously. The quantitative character of unfolding is similar for models 1 and 2, but the stability of the native state increases with the decreasing of the number of loops. In model 0, unfolding starts much more frequently from a free end of the helix, but if this process occurs at higher temperatures, it is possible to see the dissolution of a helix starting from a tight bend.

(e) *Free energy along the reaction co-ordinate*

The calculation of the free-energy change along the folding and unfolding pathways can provide substantial insight into their nature. Unfortunately, due to computational difficulties, it is impossible at present to obtain a reasonable estimate for the free energy in the transition region directly from the simulations. Thus, an alternative method for a crude estimation of the free energy is proposed. It is based on the Zimm & Bragg (1959) statistical mechanical theory of the helix–coil transition. To enable us to calculate the partition function, some simplifying assumptions are made. First, the excluded volume is neglected and second, in the denatured state, all long-range interactions are neglected. The free energy of the denatured state is given by:

$$\frac{A_D}{k_B T} = \frac{-S_D}{k_B} = -\ln Z_D, \qquad (2)$$

where Z_D is the configurational partition function of denatured state. Further details concerning the calculation of Z_D are given in the Appendix a. In the case of model 0, we neglect all entropic contributions to the free energy of the native state. This is a good approximation, because only the tails fluctuate under these conditions. Thus, the free

energy of the native A_N state is given by:

$$A_N/k_BT = E_N/k_BT, \quad (3)$$

where E_N is the configurational energy of the native state.

The estimation of free energies of the native states for models 1 and 2 is more complicated. Loops in the native structure are not completely immobile near the transition temperature, and thus their entropic contribution to the free energy must be considered. In the Appendix a, we give a prescription for estimating their entropic contribution to A_N. For these cases, the free energy of the native states is calculated according to:

$$A_N/k_BT = E_N/k_BT - \ln Z_N, \quad (4)$$

where Z_N is the configurational partition function of the loop in the native state.

The first question that must be addressed is the ability of this analytic theory to correctly reproduce the transition curves, e.g. $\langle S^2 \rangle$ as a function of T^*. The approximate transition temperature midpoint $T^*_{1/2}$, can be estimated by equating $A_N(T^*)$ and $A_D(T^*)$ and is found to be 0·667, 0·541 and 0·500 for models 01n, 1h and 2h, respectively. The transition temperature is in very good agreement with that of the simulation (Table 1) for model 0. In the remaining cases, the calculated transition mid-point is slightly lower because of the underestimate of the entropy of the native states (compare with Tables 2 and 3). The mean-square radius of gyration in the transition region is calculated and analytically according to:

$$\langle S^2 \rangle = \langle S^2_N \rangle f_N + (1 - f_N)\langle S^2_D \rangle, \quad (5)$$

where $\langle S^2_N \rangle$ and $\langle S^2_D \rangle$ are the mean-square radius of gyration of the native and denatured state, and the probability of being in the native state f_N is given by:

$$f_N = \frac{\exp(-(A_N - A_D)/k_BT)}{1 + \exp(-(A_N - A_D)/k_BT)}. \quad (6)$$

The values of $\langle S^2_N \rangle$ are taken from the simulations results and are 17·0, 19·3 and 22·4 for models 0, 1 and 2, while $\langle S^2_D \rangle$ are 65·5, 70·0 and 85·0 for models 01n, 1h and 2h, respectively. In Figure 3, in the curves denoted by the broken lines $\langle S^2 \rangle$, calculated *via* equations (8) and (9) of Paper I, is plotted *versus* temperature. One can see excellent agreement between theory and simulation of model 0, and slightly (and negligibly) shifted curves for models 1 and 2, because of the reasons outlined above.

On the basis of the good agreement between the simulations and analytic theory for the equilibrium aspects of the conformational transition, it seems reasonable to construct the free energy along the reaction co-ordinate relative to the denatured state. The free energy of a partially folded chain containing N_C contact pairs is:

$$A(N_C)/k_BT = E(N_C)/k_BT - \ln Z(N_C). \quad (7)$$

where $E(N_C)$ is the energy of the assembled part of the chain with N_C contact pairs and includes both secondary and tertiary interactions. $Z(N_C)$ is the configurational partition function of the random coil parts of the chain attached to the assembled part of the chain having N_C contact pairs (see the Appendix a for details). The most frequently observed pathway of folding is chosen to calculate the changes in free energy according to equation (7). Hence, it is assumed for model 0n1 that the hairpin consisting of helices 2 and 3 forms first with folding initiating at the bend, and then the helical turns appear according to the zipper model.. In model 1h, the chosen pathway involves the formation of the hairpin from helices 1 and 2, followed by the start of assembly of helix 3 (with the loop attached to the hydrophilic residues of the 1-2 hairpin) and, finally, the assembly of helix 4 follows. In model 2h, the hairpin composed of helices 2 and 3, forms first. Subsequently, helix 1 assembles with loop 1 interacting with the exterior of hairpin 1-2 and, finally, helix 4 is formed with loop 2 interacting with the exterior of the three-member (1-2-3 helices) bundle. The number of native contact pairs can be treated as the reaction (folding) co-ordinate, but one has to remember that this is a simplified folding model that takes into account only the steps that strictly and directly lead to the native state. All incorrectly folded species *a priori* are omitted. In Figure 11, the free energy of a model 0n1 chain having N_C native contact pairs relative to the free energy of a denatured chain $(A_{N_C} - A_D)/k_BT$ is plotted *versus* the number of native contact pairs N_C at three temperatures, $T^* = 0{\cdot}833$ (where the denatured state is preferred), at $T^* = 0{\cdot}667$ (in the

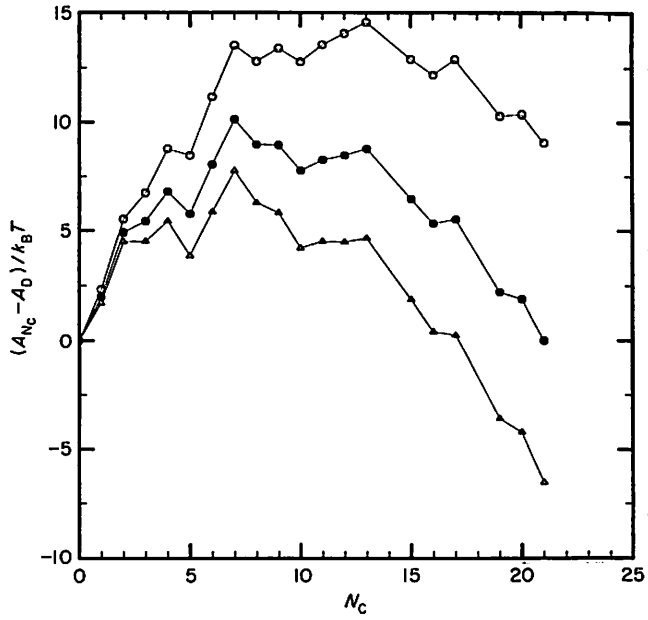


Figure 11. The plot of the relative free energy $(A_{N_C} - A_D)/k_BT$ *versus* the number of native contact pairs, N_C, for model 0n1. $T^* = 0{\cdot}833$ (open circles), 0·667 (filled circles) and 0·588 (open triangles).

transition region) and at $T^* = 0{\cdot}588$ (where the native state is preferred).

One can see from Figure 11 that at the very beginning of folding (up to $N_C = 2$) there is a steep barrier, the height of which does not change very much with temperature. This is because the process of forming a neutral bend plus one or two hydrophobic contacts is dominated by the reduction in entropy, and the dependence of entropy on temperature is weak. The next steps (up to $N_C = 4$) correspond to the completion of folding of the entire 2-3 hairpin. The free energy of the chain increases during this stage, but this increase is less than up to $N_C = 2$ because of the additional hydrophobic contacts. The height of the barrier at $N_C = 2$ takes the value of 8·76 at $T^* = 0{\cdot}833$, 6·80 at $T^* = 0{\cdot}667$ and 5·42 at $T^* = 0{\cdot}588$. Finally, for $N_C = 5$, the energy decrease overrides the loss of entropy and a local minimum in the free energy appears. The free energy at this point decreases with increasing temperature and assumes the values 8·46 at $T^* = 0{\cdot}833$, 5·76 at $T^* = 0{\cdot}667$ and 3·84 at $T^* = 0{\cdot}588$. The presence of this minimum is the reason why relatively stable hairpins are present even at temperatures higher than the folding transition; moreover, this is the most stable intermediate state in the vicinity of the transition. This differs from the folding of the Greek key, β-motif, where a single β-hairpin is not stable (Skolnick & Kolinski, 1989). The following steps (up to $N_C = 7$) are very similar to the first ($N_C \geqslant 2$) and correspond to the formation of the bend between helices 2 and 1 and the initiation of helix 1, a process that is entropically controlled. The next step (up to $N_C = 10$, where helix 1 is completed) leads to a local minimum of the free energy that corresponds to a three-helix bundle. The difference of free energy between this state and the denatured state changes strongly with the temperature and is 12·75 at $T^* = 0{\cdot}833$, 7·76 at $T^* = 0{\cdot}667$ and 4·20 at $T^* = 0{\cdot}588$. Between the three-helix bundle and the native state there is a small barrier of entropic origin, but the assembly of this last helix (up to $N_C = 21$) produces twice as many native contacts per step as for the previous ones. The entropic barrier between $N_C = 10$ and $N_C = 13$ is very small, and equals 1·81 at $T^* = 0{\cdot}833$, 1·01 at $T^* = 0{\cdot}667$ and 0·44 at $T^* = 0{\cdot}588$. The transition state at $N_C = 13$ consists of the three-member bundle plus the native bend and three-quarters of the first α-helical turn in the fourth helix. Hence, this state is close to the native state in both configuration and entropy, and its energy differs from that of the native state by $\varepsilon_h + 8\varepsilon_c$. The remaining steps are characterized by a large decrease in energy. The native state is always in at least a local free-energy minimum and, at lower temperatures, this is the global minimum as well. The barrier height with respect to the intermediate is much less dependent on the temperature than with respect to the native state. This implies that the rate of unfolding decreases more rapidly with the change in conditions than that of refolding, in agreement with experimental results (Brems *et al.*, 1982; Tsong & Baldwin, 1978).

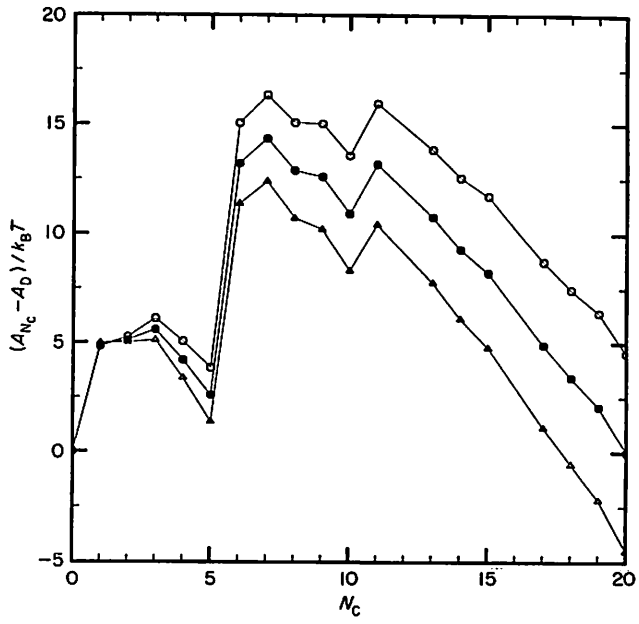


Figure 12. A plot of the relative free energy $(A_{N_C} - A_D)/k_BT$ *versus* the number of native contact pairs, N_C, for model 1h. $T^* = 0{\cdot}588$ (open circles), 0·541 (filled circles) and 0·500 (open triangles).

In Figure 12, the relative free energy $(A_{N_C} - A_D)/k_BT$ *versus* the number of native compact pairs, N_C, is plotted for a representative folding pathway for model 1h. The temperatures 0·588, 0·541 and 0·500 are chosen to be representative of the denatured state, the transition region and the native state, respectively. The change of the free energy due to the assembly of the first hairpin is virtually the same as in model 0. But it should be pointed out that the transition temperature is much lower for this model, and thus both intermediate states (hairpins and 3-member bundles) are more stable. The difference of free energy between the denatured state and a hairpin is 3·85 at $T^* = 0{\cdot}588$, 2·59 at $T^* = 0{\cdot}541$ and 1·37 at $T^* = 0{\cdot}500$. The main difference between models 0 and 1 lies in the huge free-energy barrier between the first hairpin and the beginning of the third helix (between $N_C = 5$ and 6). This is a predominantly entropic barrier, which is caused by freezing of this loop from a random coil with one end pinned to only a few, very similar, configurations. The dominance of the entropic contribution produces a barrier that is almost independent of temperature; the free-energy difference between these two states is 11·18 at $T^* = 0{\cdot}588$, 10·59 at $T^* = 0{\cdot}541$ and 9·88 at $T^* = 0{\cdot}500$. The assembly of the entire third helix ($N_C = 10$) leads to a local free-energy minimum, which is deeper than that for model 0, because of the lower transition temperature. The relative free energy of a three-member bundle is very high; 13·55 at $T^* = 0{\cdot}588$, 10·91 at $T^* = 0{\cdot}541$ and 8·31 at $T^* = 0{\cdot}500$. The assembly of the last helix is very

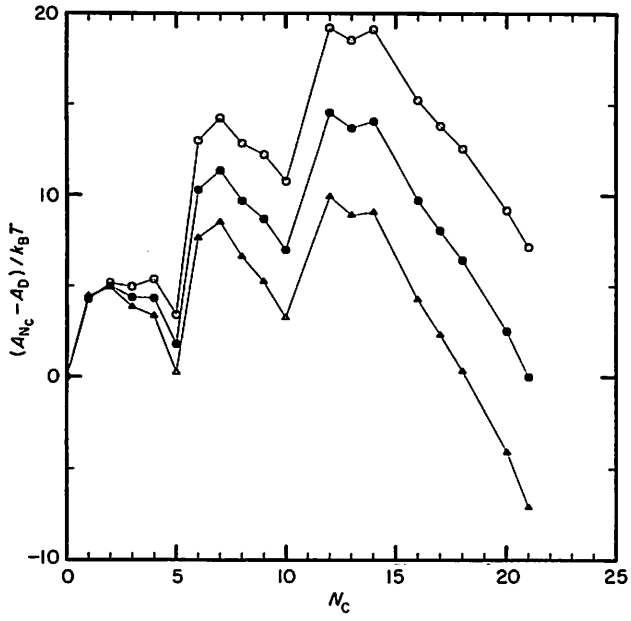


Figure 13. A plot of the relative free energy $(A_{N_C}-A_D)/k_BT$ *versus* the number of native contact pairs, N_C, for model 2h. $T^* = 0.556$ (open circles), 0·500 (filled circles) and 0·455 (open triangles).

similar to that in model 0. There is a relatively small entropic barrier between the three-member bundle and the native state, located at $N_C = 11$. The height of this barrier is 2·39 at $T^* = 0.588$, 2·25 at $T^* = 0.541$ and 2·12 at $T^* = 0.500$. The transition state at $N_C = 11$, corresponding to this barrier, consists of a three-helix bundle plus the native bend and three-quarters of the first α-helical turn in the fourth helix. The energy difference between the native state and the transition state is $9\varepsilon_h + 8\varepsilon_c$.

It should be pointed out that a single hairpin (1-2 or 3-4) is a very stable intermediate state and has a free energy comparable with that of denatured state (for $T^* \geqslant T^*_{1/2}$), thus it is a rather well-populated state. This is confirmed by the direct simulation results in Table 2. The three-member bundle intermediate has a higher energy and is much less populated over the entire simulation run.

In Figure 13, for model 2h, the relative free energy $(A_{N_C}-A_D)/k_BT$ is plotted *versus* the number of native contacts N_C at temperatures 0·556 (denatured), 0·500 (transition) and 0·455 (native). The changes of free energy during the assembly of the first (2-3) hairpin are very close to those in the previous case. The free energy of this hairpin is low; 3·40 at $T^* = 0.556$, 1·79 at $T^* = 0.500$ and 0·24 at $T^* = 0.455$. Thus, at lower temperatures the hairpin should be populated equally with the denatured state, which is in good agreement with direct simulation results (Table 3). The next intermediate state, the three-member bundle, has a higher free energy (10·76 at $T^* = 0.556$, 6·96 at $T^* = 0.500$ and 2·34 at $T^* = 0.455$). Between the first hairpin and the three helix bundle, the configuration of the long loop L_1 is almost frozen, and consequently there is a huge entropic barrier as in the previous case. The height of this barrier is 10·83 at $T^* = 0.556$, 9·56 at $T^* = 0.500$ and 8·28 at $T^* = 0.455$. The second entropic barrier between $N_C = 10$ and $N_C = 12$, caused by freezing of the second loop, has a height of 8·24 at $T^* = 0.556$, 7·57 at $T^* = 0.500$ and 6·68 at $T^* = 0.455$. Both these barriers are weakly dependent on temperature because of their entropic nature. But contrary to models 0 and 1, the three-member helical bundle is located in a relatively deep local minimum surrounded by barriers on the left and right. This means that this state becomes more populated over the entire Monte Carlo run, and when a successful folding event takes place, the model chain is in this state for longer periods of time than in the hairpin conformation. The transition state corresponding to the barrier at $N_C = 12$ consists of a three-member bundle with the second loop frozen on the external surface of the bundle plus half of the first α-helical turn of the fourth helix. The energy difference between the native state and the transition state is $9\varepsilon_h + 11\varepsilon_c$.

4. Summary and Conclusion

Here, we have examined the folding and unfolding pathways of α-helical hairpins and all variants of left-handed, four-helix bundles. The folding of an α-helical hairpin is designed to explore whether the diffusion-collision-adhesion model, wherein isolated α-helices form first and then diffuse together to produce the hairpin (Karplus & Weaver, 1976, 1979; Lee *et al.*, 1987) is correct. While this mechanism is a competitive possibility, in fact, folding was seen to initiate at or near a native turn, accompanied by the formation of a hydrophobic cluster, followed by growth of both helices on-site. In general, whether or not the on-site construction mechanism dominates over a diffusion-collision-adhesion mechanism depends on the relative ratio of the mean life-time of isolated α-helices compared to the mean time of their coalescence. If the isolated helices live sufficiently long relative to the time for successful coalescence, then the diffusion-collision-adhesion mechanism of assembly dominates the folding process. However, for cases where the isolated α-helices are marginally stable (for example, here, where the model in the denatured state has a mean helix content of 20 to 30%), it is faster and more probable to dissolve them and construct the hairpin on-site. Thus, the latter mechanism dominates for the physically reasonable range of helix stabilities.

Turning to the folding pathway of all variants of four helix bundles, two folding intermediate states are found. The first is an α-helical hairpin, and the second is identified to consist of three of the four native helices assembled. Assembly initiates from the formation of an α-helical hairpin to which the next helical stretch attaches. The hairpin state is separated from the denatured state by a predominantly energetic barrier and is separated from the second intermediate by an entropic barrier. The

first barrier is virtually the same for all models under consideration, and the second barrier is much higher for models with a long loop(s). The three-helix bundle is separated from the native state by a free-energy barrier of mainly entropic origin. This barrier is relatively small in the case of models with only tight bends and with one long loop. The free energy of the denatured state and that of the α-helical hairpin are comparable in the transition region and, thus, the conformation of the model chain oscillates between the randomly coiled state and a hairpin with coiled tail(s). The only exception to this is the model with tight bends; here, the α-helical hairpin is not so stable because the transition occurs at a higher temperature. After the assembly of an intermediate state, the remaining random coil tail(s) thrashes about searching for the relatively narrow pass in configuration space to the native state. The excluded volume effect exerted by already assembled tertiary structure aids in assembly in the early stages and (especially in the case of models containing long loops where the helix may have to "snake" under the loop in order for the native state to form) hinders assembly in the later stages.

The dominant mechanism of assembly for the four-helix bundle with tight bends, consisting of sequential assembly of a pair of adjacent hairpins, followed by attachment of the adjacent third and then fourth helices, differs from the folding pathways of cytochrome *c* deduced from hydrogen-exchange labeling and proton nuclear magnetic resonance spectroscopy by Roder *et al.* (1988), where the N and C-terminal helices assemble first. While we have seen such events, they are very rare; the reduction in configurational entropy required to form the N and C-terminal helical eyelet with a long intervening random coil loop between them is very large. Perhaps these folding initiation events could be made more frequent by increasing the stability of the N and C-terminal helices (all the helices in these simulations are equally stable.) More likely, this reflects the influence of the heme group on the folding pathway, an effect that is entirely absent in our model. In forthcoming simulations, we shall attempt to identify conditions in the model that reproduce this folding pathway. To clarify the experimental situation, hydrogen-exchange studies should be done on the synthetic, four-helix bundles of uniform stability synthesized by Regan & DeGrado (1988).

While a general trend of folding has been observed, there are (especially in the early stages) a multiplicity of pathways that lead to the early intermediates. As the folding progresses further, while the number of options becomes more limited in a general sense, there remains many spatial trajectories consistent with the overall pathway of assembly. For example, after the three-helix bundle intermediate has formed, the random coil tail depending on chance, may be on the correct side of the protein to fold, or on the wrong side; in which case, it must work its way around to the correct side before assembly can occur. All the above events can be described as thrashing about of the tail before final assembly, but the details differ from one folding event to the other.

We have also performed, using an approximate analytic theory, an analysis of the free energy along the dominant folding pathways. The transition state is identified for all cases as an almost fully assembled, but not distorted native conformation; this is to be contrasted with the "Cardboard Box" model proposed by Goldenberg and Creighton (1985), which conjectures that the transition state is a distorted native conformation. The free-energy barrier relative with respect to the intermediate is primarily entropic in origin, with the transition state being very close in entropy to the native state; the free-energy difference between the transition state and the native state is predominantly energetic. In this, the mechanism of assembly agrees with the Cardboard Box model, which also conjectures that the native state is kinetically trapped. Qualitatively, the simulations agree with experiment in both the relative temperature dependence of folding and unfolding (Brems *et al.*, 1982; Tsong & Baldwin, 1978) with the latter being more sensitive to conditions, and the fact that the transition state lies close to the native state.

In conclusion, while this is a highly simplified model of α-helical proteins, nevertheless it exhibits a remarkable number of features of the equilibrium and dynamic aspects of the globular protein folding/unfolding conformational transition. Since the on-site construction mechanism obtains for both α-helical and β-protein models, we believe it to be quite general, and this qualitative description of folding should be applicable to real proteins. Nevertheless, further refinement of this class of models is required to make them more realistic and less idealized; such generalization are in progress.

APPENDIX

Calculation of the Configurational Partition Function of a Modified Zimm–Bragg Model

In our simple analytical model, we consider a chain consisting of n residues having $n-3$ rotational degrees of freedom. A rotational state specifies whether a given residue is in a helical (h) or coiled (c) state. A helical state corresponds to a g^- state and a coiled state to g^+ or t states. The kth conformational state is formed by the (k)th, $(k+1)$th, $(k+2)$th and $(k+3)$th residues, and hence

conformational states are defined for residues 1, ..., $(n-3)$. The coupling of the conformational states occurs due to the helical wheel-type interactions and involves five consecutive residues: residue i can interact with residue $i+4$, and thus the number of these coupled states is $n-4$. In our notation, the jth helical wheel interaction involves residues from j to $j+4$. A given conformational state can participate in different helical wheel interactions; the 19th helical wheel interaction (between residues 19 and 23) involves the 19th and 20th rotational states, and the adjacent helical wheel interaction number 20 (between residues 20 and 24) involves the 20th and 21st rotational states. In order to make the analytical model tractable, we assume that a denatured chain has no excluded volume and that tertiary interactions are absent. This means that the many-body problem is reduced to a nearest neighbor, one-dimensional Ising model (Poland & Scheraga, 1970), and hence it is easily soluble. This simplification enables us to calculate the thermodynamic functions of a purely denatured state taking into account only the entropic and local energetic contributions. For partially folded structures, this implies that the entropic and tertiary energetic contributions can be calculated separately.

In order to calculate the configurational partition function of the denatured state, we have to construct statistical weights matrices for adjacent states interacting with a helical wheel potential ε_c:

$T_i =$

$^{2i+1,2i+2}_{2i-1,2i}$	cc	ch	hc	hh
cc	4	2	2	w
ch	4	2	$2w$	w^2
hc	4	2	2	w
hh	4	2	$2w$	w^2

(A1)

where $w = \exp(-\varepsilon_C/k_BT)$, and i runs from 0 to $(n-2)/2$. It is assumed that the statistical weights of coiled states are equal to 1. The first column, being devoid of helical wheel interactions should contain only fours; similarly, the second and the third column should contain only twos. The appearance of any new helical wheel interaction introduces an additional factor w. Four the first and the last states in a chain, special matrices are constructed. A dummy coiled state should be added before the first and after the last state. If the number of rotation states is odd, then the matrix associated with rotational state $n-3$ is:

$\frac{T_{n-2}}{2} =$

	cc	ch	hc	hh
cc	2	0	1	0
ch	2	0	w	0
hc	2	0	1	0
hh	2	0	w	0

(A2)

In the case of a neutral bend/loop region, matrix (A1) is slightly modified by setting $w = 1$ for the appropriate states. According to the classical treatment by Flory (1969), the configurational partition function of a denatured chain is given by:

$$Z_D = (1,0,0,0) \prod_{i=0}^{(n_h)} T_i(1,1,1,1)^T \qquad \text{(A3)}$$

where $n_h = (n-3)/2$ if n is even, and $n_h = (n-4)/2+1$ if n is odd.

All the above considerations deal with a model chain without tertiary interactions, and hence can be applied to the high-temperature denatured chain, where the number of nearest-neighbor contacts is very small and can be neglected. In the case of a partially folded chain, we assume that it is possible to consider the contributions to the free energy of the tertiary interactions and the short-range interactions separately. The total free energy of the folded chain, calculated from equation (10) of Paper I, is divided into two parts. The energetic contribution, E_{N_C}, includes tertiary and secondary interactions of a partially assembled chain having N_C contact pairs. $Z(N_C)$ is the configurational partition function of the remainder of the chain that is devoid of native contacts. We assume that the statistical weight matrices of these unfolded parts of the chain remain unaltered on comparison with the denatured state and are thus given by equation (A1) or (A2), where appropriate. The folded part of a chain has different matrices. For example, the allowed conformations of the pair of residues at the start of the native helical stretch are of the form:

$T_i =$

	cc	ch	hc	hh
cc	0	2	0	w
ch	0	2	0	0
hc	0	2	0	w
hh	0	2	0	0

(A4)

that is, we assume that the residue immediately preceding the native conformation is randomly coiled. Furthermore, the matrices corresponding to rotational states inside a folded helix have the following form:

	cc	ch	hc	hh
cc	0	0	0	0
ch	0	0	0	0
hc	0	0	0	0
hh	0	0	0	1

(A5)

because only hhhh sequences exist in a helix (with statistical weight 1). Finally, the matrix associated with interfacial residues to the right of the assembled tertiary structure is of the form:

$T_i =$

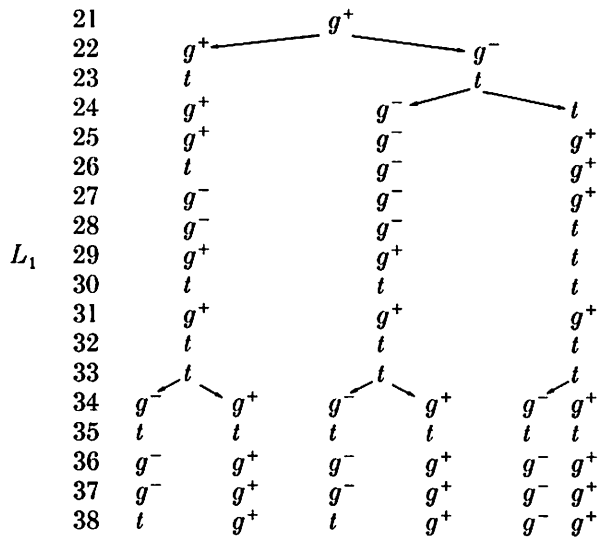
	cc	ch	hc	hh
cc	0	0	0	0
ch	0	0	0	0
hc	4	2	2	w
hh	4	2	0	0

(A6)

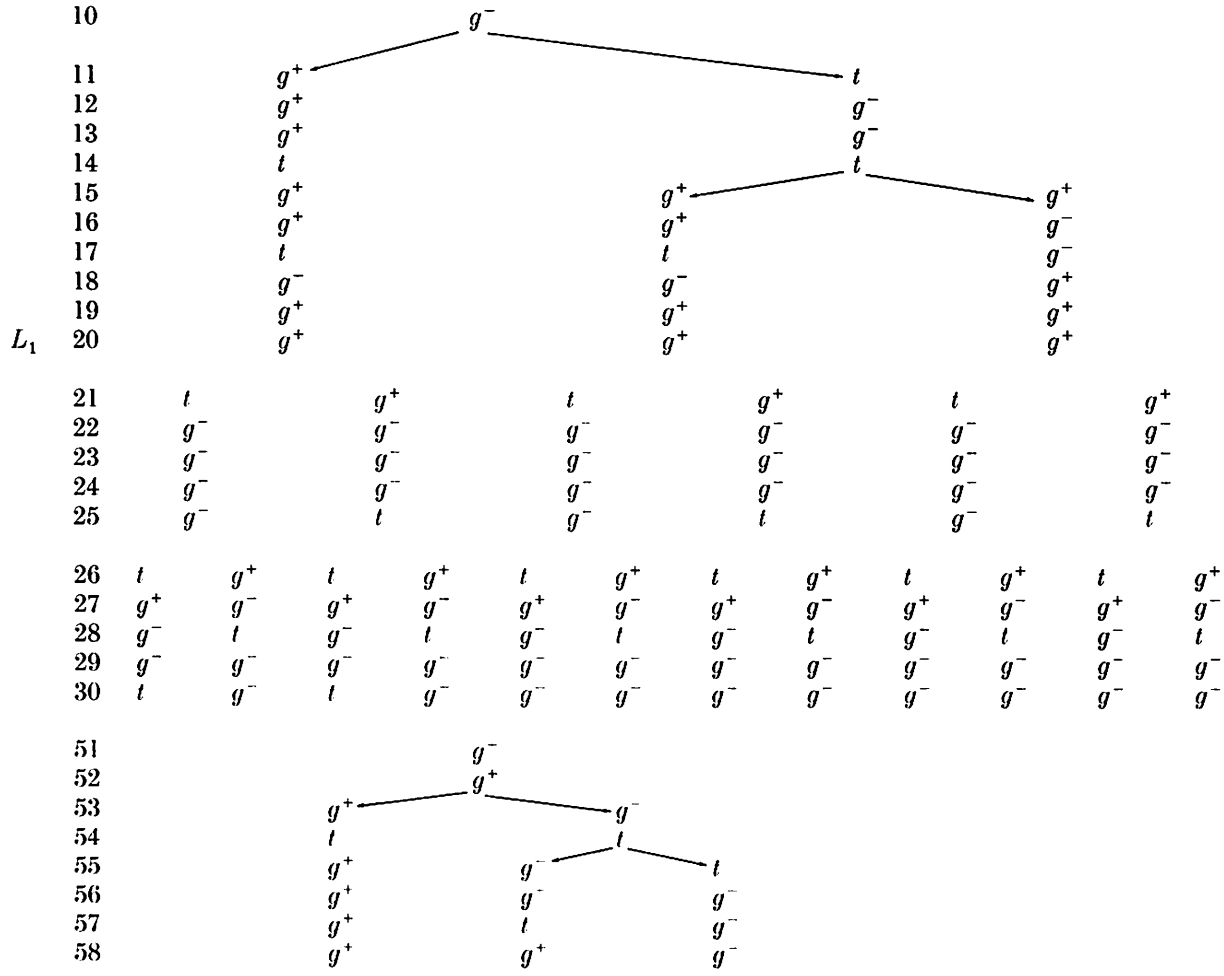
For models 1 and 2, the long loop(s) can be in a purely random state (see above) or can be in the "native" state connecting native helices. In the latter case, it is assumed that the loop has limited mobility and that the number of energetic contacts with the hydrophilic exterior of the already assembled structure is constant. This number is taken from the simulation. There are five contacts in model 1 between loop L_1 and hairpin 1-2. In model 2, there are eight contacts between loop L_1 and hairpin 2-3, and there are five contacts between loop L_1 and the three-member bundle 1-2-3. Loops can locally change their conformation between contacts and can also change conformation by shifting contacts over short distances (2 to 3 segments). This behavior is actually observed in the simulations at lower temperatures. Hence, the entropic contribution to the free energy is non-zero even in the native state where loops are present.

Here, we explicitly list the rotational states of the loops that are allowed in the native states and that are used in statistical weight matrices by assigning 1 to the state indicated below and zero for all other conformations of the residue.

Model 1



Model 2



		Model 2		
Residue no.				
L_2	59	t	t	t
	60	g^-	g^-	g^-
	61	t	t	t
	62	t	t	t
	63	g^-	g^-	g^-
	64	t	t	t
	65	g^-	g^-	g^-
	66	g^+	g^+	g^+
	67	t	t	t

This work was supported in part by NIH grant GM37408 from the Division of General Medical Sciences, United States Public Health Service. Stimulating conversations with Drs David Case, Alfred Holtzer, Andrzej Kolinski and Peter Wright are gratefully acknowledged.

References

Abdel-Meguid, S. S., Shieh, H. S., Smith, W. W., Dayringer, H. E., Violand, B. N. & Bentle, L. A. (1987). *Proc. Nat. Acad. Sci., U.S.A.* **84**, 6434–6437.
Anfinsen, (1972).
Barlow, D. J. & Thornton, J. M. (1988). *J. Mol. Biol.* **201**, 601–619.
Baumgartner, A. (1984). *Annu. Rev. Phys. Chem.* **35**, 419–435.
Binder, K. (1987). Editor of *Monte Carlo Methods in Statistical Physics*, chapt. 1. Springer, Berlin.
Brems, D. N., Cass, R. & Stellwagen, E. (1982). *Biochemistry*, **21**, 1488–1493.
Chandrasekhar, S. (1943). *Rev. Mod. Phys.* **15**, 1–89.
Creighton, T. E. (1974). *J. Mol. Biol.* **87**, 563–577; 603–624.
Creighton, T. E. (1977*a*). *J. Mol. Biol.* **113**, 275–293.
Creighton, T. E. (1977*b*) *J. Mol. Biol.* **113**, 329–341.
Creighton, T. E. (1979). *J. Mol. Biol.* **129**, 411–431.
Creighton, T. E. (1985). *J. Phys. Chem.* **89**, 2452–2459.
Creighton, T. E. (1988). *Proc. Nat. Acad. Sci., U.S.A.* **85**, 5082–5086.
Dyson, H. J. Rance, M., Houghton, R. A., Lerner, R. A. & Wright, P. E. (1988*a*). *J. Mol. Biol.* **201**, 161–200.
Dyson, H. J., Rance, M., Houghten, R. A., Lerner, R. A. & Wright, P. E. (1988*b*). *J. Mol. Biol.* **201**, 201–217.
Flory, P. J. (1956). *J. Amer. Chem. Soc.* **78**, 5222–5235.
Ford, G. C., Harrison, P. M. Rice, D. W., Smith, J. M. A., Treffry, A., White, J. L. & Yariv, J. (1984). *Phil. Trans. Roy. Soc. ser. B*, **304**. 551–565.
Garel, J. R. & Baldwin, R. L. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**. 3347–3351.
Ghelis, & Yon, (1982).
Gō, N., Abe, H., Mizuno, H. & Taketomi, H. (1980). *Protein Folding* (Jaenicke, N., ed.), pp. 167–181. Elsevier/North Holland. Amsterdam.
Goldenberg, D. P. & Creighton, T. E. (1985). *Biopolymers*, **24**, 167–182.
Guss, J. M. & Freeman, H. C. (1983). *J. Mol. Biol.* **169**, 521–563.
Jacobson, H. & Stockmayer, W. H. (1950). *J. Chem. Phys.* **18**. 1600–1606.
Janin, J. & Chothia, C. (1980). *J. Mol. Biol.* **143**, 95–128.
Karplus, M. & Weaver, D. L. (1976).
Karplus, M. & Weaver, D. L. (1979). *Biopolymers*, **18**, 1421–1427.
Kim, & Baldwin, R. L. (1982).
Kolinski, A., Skolnick, J. & Yaris, R. (1986*a*). *J. Chem. Phys.* **85**, 3585–3597.
Kolinski, A., Skolnick, J. & Yaris, R. (1986*b*). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 7267–7271.
Kolinski, A., Skolnick, J. & Yaris, R. (1987). *Biopolymers*, **26**, 937–962.
Lee, (1987).
Matheson, R. R. Scheraga, H. A. (1978). *Macromolecules*, **11**, 819–829.
Privalov, P. L. (1979). *Advan. Protein Chem.* **33**, 167–241.
Poland, D. & Scheraga, H. A. (1970). *Theory of Helix-Coil Transitions in Biopolymers*, Academic Press, New York.
Regan, L., DeGrado, W. F. (1988). *Science*, **241**, 976–978.
Richardson, J. S. (1981). *Advan. Protein Chem.* **34**, 167–339.
Roder, H., Elöve, G. A. & Englander, W. (1988). *Nature (London)*, **335**, 700–704.
Schiffer, M. D. & Edmundson, A. (1967). *Biophys. J.* **1**, 121–135.
Sikorski, A. & Skolnick, J. (1989*a*). *Biopolymers*, **28**, 1097–1113.
Sikorski, A. & Skolnick, J. (1989*b*). *Proc. Nat. Acad. Sci., U.S.A.* **86**, 2668–2672.
Skolnick, J. & Kolinski, A. (1989). *J. Mol. Biol.*
Skolnick, J., Kolinski, A. & Yaris, R. (1988). *Proc. Nat. Acad. Sci., U.S.A.* **85**, 5057–5061.
Skolnick, J., Kolinski, A. & Yaris, R. (1989*a*). *Biopolymers*, **28**, 1059–1095.
Skolnick, J., Kolinski, A. & Yaris, R. (1989*b*). *Proc. Nat. Acad. Sci., U.S.A.* **86**, 1229–1233.
Taketomi, H., Kano, F. & Gō, N. (1988). *Biopolymers*, **27**, 527–559.
Tanford, (1968).
Tsong, T. Y. & Baldwin, R. L. (1978). *Biopolymers*, **17**, 1669–1678.
Udgaonkar, J. B. & Baldwin, R. L. (1988). *Nature (London)*, **335**, 694–699.
Weber, P. C. & Salemme, F. R. (1980). *Nature (London)*, **287**, 82–84.
Wetlaufer, D. B. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 697–701.
Wright, (1988).
Zimm, & Bragg, (1959).

Edited by T. J. Richmond