Reprinted from *J. Mol. Biol.* (1989) 212, 787–817

Dynamic Monte Carlo Simulations of Globular Protein Folding/Unfolding Pathways

I. Six-member, Greek Key β-Barrel Proteins

Jeffrey Skolnick and Andrzej Kolinski

J. Mol. Biol. (1989) 212, 787–817

Dynamic Monte Carlo Simulations of Globular Protein Folding/Unfolding Pathways

I. Six-member, Greek Key β-Barrel Proteins

Jeffrey Skolnick†

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

and Andrzej Kolinski

Department of Chemistry
University of Warsaw, 02-093 Warsaw, Poland

(Received 4 April 1989; accepted 26 September 1989)

In the context of a simplified diamond lattice model of a six-member, Greek key β-barrel protein that is closely related in topology to plastocyanin, the nature of the folding and unfolding pathways have been investigated using dynamic Monte Carlo techniques. The mechanism of Greek key assembly is best described as punctuated "on site construction". Folding typically starts at or near a β-turn, and then the β-strands sequentially form by using existing folded structure as a scaffold onto which subsequent tertiary structure assembles. On average, β-strands tend to zip up from one tight bend to the next. After the four-member, β-barrel assembles, there is a long pause as the random coil portion of the chain containing the long loop thrashes about trying to find the native state. Thus, there is an entropic barrier that must be surmounted. However, while a given piece of the protein may be folding, another section may be unfolding. A competition therefore exists to assemble a fairly stable intermediate before it dissolves. Folding may initiate at any of the tight turns, but the turn closer to the N terminus seems to be preferred due to well-known excluded volume effects. When the protein first starts to fold, there are a multiplicity of folding pathways, but the number of options is reduced as the system gets closer to the native state. In the early stages, the excluded volume effect exerted by the already assembled protein helps subsequent assembly. Then, near the native conformation, the folded parts reduce the accessible conformational space available to the remaining unfolded sections. Unfolding essentially occurs in reverse. Employing a simple statistical mechanical theory, the configurational free energy along the reaction co-ordinate for this model has been constructed. The free energy surface, in agreement with the simulations, provides the following predictions. The transition state is quite near the native state, and consists of five of the six β-strands being fully assembled, with the remaining long loop plus sixth β-strand in place, but only partially assembled. It is separated from the β-barrel intermediate by a free energy barrier of mainly entropic origin and from the native state by a barrier that is primarily energetic in origin. The latter feature is in agreement with the "Cardboard Box" model described by Goldenberg and Creighton but, unlike their model, the transition state is not a high-energy distorted form of the native state. The theory predicts that the rate of folding is less sensitive to changes in folding conditions than is the rate of unfolding, in agreement with experiment. Finally, the simple theory provides a means of assessing the effects of amino acid substitutions that favor native-like turn formation. By stabilizing the intermediate, they enhance the rate of folding to the intermediate from the denatured state

† Permanent address: Department of Molecular Biology, Research Institute of Scripps Clinic, 10666 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.

0022-2836/90/080757-31 $03.00/0

and the rate of unfolding out of the native state, and they act to slow down the rate of folding from the intermediate to the native state. But even when every turn residue is modified, the effect is relatively small.

1. Introduction

The elucidation of the mechanism by which a globular protein finds its way to native conformation from the denatured state is one of the most tantalizing. unsolved problems in contemporary molecular biololgy (Kim & Baldwin, 1982; Anfinsen, 1972; Harrison & Durbin, 1985; Creighton, 1985, 1988). It has long been recognized that the mean time required for a random search through all of configuration space to find the native state far exceeds the observed folding times (Levinthal, 1968; Wetlaufer, 1973; Karplus & Weaver, 1976), which are of the order of seconds or minutes (Creighton, 1985; Garel & Baldwin, 1973; Kawajima *et al.*, 1985); thus, proteins must possess a means of partitioning configuration space so that only relevant regions are sampled. Over the years, there have been a number of suggestions how such partitioning might occur, these will be summarized below (Anfinsen, 1972; Wetlaufer, 1973; Ptitsyn & Rashin, 1975; Ptitsyn & Finkelstein, 1980). Whatever the actual mechanism may be, it is unfortunately not possible experimentally to follow directly the folding process from the nascent state all the way to the fully folded protein i.e. to make what is effectively a "movie" of the folding/unfolding pathway(s). Consequently, computer simulations may prove helpful in elucidating some of the qualitative features of protein folding. Thus, we have embarked on a series of simulations designed to provide qualitative insight into the folding and unfolding processes. This paper describes the results from dynamic Monte Carlo (MC†) simulations of a model six-stranded, Greek key, β-barrel protein (Richardson, 1981) having a topology quite close to that of plastocyanin (Guss & Freeman, 1983) and for which the pathways of folding and unfolding have been determined under *in vitro* conditions. The accompanying paper presents a similar analysis for the folding and unfolding pathways of the left-handed, four-helix bundles with tight bends, and with one and two long loops (Sikorski & Skolnick, 1990).

Before summarizing the essential features of extant folding models, it is important to review the experimental facts that any successful model must be consistent with. It is convenient to divide the folding process into two stages. Those events associated with the initiation of folding and those events involved in the latter stages when the protein is near the native state. The nature of the early folding events is considered first. Under denaturing conditions, hydrophobic clusters (Bundi *et al.*, 1976, 1978), nascent helices (Shoemaker *et al.*, 1985, 1987; Dyson *et al.*, 1988*a*,*b*), reverse β-turns (Dyson *et al.*, 1988*a*), etc. have been observed. Furthermore, small protein peptide fragments have been shown to adopt secondary structures in rapid equilibrium with unfolded species. Thus, the denatured state is not a pure random coil that is entirely devoid of any secondary structure. Oas & Kim (1988) have demonstrated the stability of a 30 residue synthetic analog of the first folding intermediate in the folding pathway of bovine pancreatic trypsin inhibitor (BPTI). Similarly, Udgaonkar & Baldwin (1988) have provided evidence from nuclear magnetic resonance studies, of an early folding intermediate in ribonuclease A. While these observations are basically equilibrium in nature, it is not unreasonable to suppose that these nascent structures have something to do with protein folding, in that they automatically provide a means of subdividing configuration space.

A particularly important observation of protein folding kinetics is that the rate-determining step seems to occur near the end of the folding pathway (Creighton, 1985, 1988). Consequently, nucleation models (Wetlaufer, 1973), which assume that once initiation of folding occurs the reaction co-ordinate is downhill in free energy (as in the growth of a liquid drop from the vapor phase) appear to be incorrect. Moreover, on the basis of the effect of crosslinks in BPTI on the folding kinetics, the transition state has been conjectured to be native-like in many, but not all, respects (Creighton, 1985, 1988). This viewpoint has been advanced by Goldenberg & Creighton (1985). In particular, the kinetics of the unfolding process in proteins is almost always well approximated by an all-or-none model (Creighton, 1988; Brandts *et al.*, 1975). Namely, a single rate constant adequately describes the unfolding kinetics, and intermediates have proved difficult to detect. These observations lend credence to the viewpoint that the transition state lies closer to the folded than the unfolded conformation (Creighton, 1981).

A number of models for protein folding have been proposed. The framework models envisage folding as proceeding along a limited number of well-defined sequential pathways having well-defined intermediates. In its simplest form, one assumes the secondary structure (e.g. an α-helix) present in the native state forms first. This is followed by the coalescence of the secondary structural elements that produce the native conformation. Karplus and co-workers (Karplus & Weaver, 1976, 1979; Lee *et al.*, 1987) have developed a series of diffusion-collision-adhesion models based on this concept. One

† Abbreviations used: MC, Monte Carlo; BPTI, bovine pancreatic trypsin inhibitor; N, native; D, denatured; I, intermediate.

possible realization of the model involves assembly *via* "prefabricated construction". The individual building blocks (e.g. α-helices) are assumed to be sufficiently stable without tertiary interactions that they have time to diffuse together before dissolution occurs. Among other things, the appeal of this model lies in its simplicity. Rather than having to worry about tertiary interactions, one need only consider short-range interactions (down the protein chain) to predict the resulting structural elements of the native state. Moreover, folding in the model is essentially sequential, and the protein does not have to adopt irrelevant conformations for subsequent native state formation.

There are a number of potential problems associated with the prefabricated construction realization of the framework model. First, there is the rather marginal intrinsic stability of secondary structures in the absence of tertiary interactions. In fact, β-strands require interstrand stabilization to impart substantial stability, but isolated α-helices might be sufficiently stable. Thus, diffusion between performed secondary structural elements (microdomains) must occur before these elements have time to dissolve. The mutual approach of two such microdomains must involve the dragging of at least a portion of the attached random coil tails, with a corresponding diminution in the relative rate of approach. As pointed out by Lee *et al.* (1987), inclusion of hydrodynamic interactions also serves to slow down their mutual approach. Furthermore, in this model it is not at all obvious why the transition state lies closer to the native rather than to the denatured state. We return below to a further analysis of the implications of the framework model. Secondly, it is known that certain short sequences of amino acids in globular proteins may be found in more than one structural motif (Zielenkiewicz & Rabczenko, 1988). At the very least, this implies that some structural arrangements due to tertiary interactions must also occur.

An alternative model involves modular assembly (Kim & Baldwin, 1982; Harrison & Durbin, 1985), wherein subdomains are assumed to fold without any pre-existing secondary structure. This model addresses the marginal stability of secondary structure without tertiary interactions by asserting that tertiary interactions alone are responsible for secondary structure. Such a viewpoint is not inconsistent with the thermodynamics of the conformational transition (Privalov, 1979) but, if the model is taken literally, it is not clear how the system can prepartition configuration space in the initial stages of protein folding. Moreover, there is marginal secondary structure in the denatured state. There are also some theoretical analyses that indicate that systems lacking marginal, intrinsic secondary structure stability should collapse to very dense, nonunique randomly coiled globules, with patches of secondary structure, rather than the ordered structures observed in globular proteins (Kolinski *et al.*, 1987*a*).

Wright *et al.* (1988) proposed a model of single-domain protein assembly that is an amalgamation of elements of both the framework and modular assembly models. On the basis of evidence from nuclear magnetic resonance studies, they strongly and convincingly argue that the formation of nascent helical structures, β-turns and hydrophobic clusters may be the essential events in early protein folding. However, a crucial observation is that these structures are marginally stable and in rapid equilibrium with unfolded states. They further argue that these transient structures then diffuse together and result in structures of enhanced, but still marginal, stability. Further progression along the folding pathway produces a globular polypeptide having a hydrophobic core with extensive secondary and supersecondary structure. This corresponds to the first observable intermediate. This structure then rearranges into a compact, folded but nonnative structure. Finally, and most often the rate-determining step, the rearrangement of the nonnative structure into the native conformation occurs. The latter is similar to the "Cardboard Box" model proposed by Goldenberg & Creighton (1985).

Clearly, given the broad disparity in the models described above, a very large number of questions concerning the nature of the folding process are unanswered at this time. Computer simulations of the folding process can, in principle, provide a number of insights into the nature of the folding process. They can provide a full trajectory. Consider, for example, the formation of an α-helical hairpin. Even if one has established that the isolated helices are marginally stable and that the hairpin exists, it does not automatically follow that the isolated helices remained stable long enough for them to diffuse together and form a hairpin. Other alternatives include the possibility of hairpin formation due to side by side growth of both helices from a hydrophobic cluster or the formation from a single helix and a turn onto which the second helix was constructed on site. Unfortunately, knowledge of the initial and final conformations without knowledge of the time-course of the intervening processes does not allow one to specify the reaction co-ordinate. Since computer simulations can provide trajectories, they are a particularly powerful tool for elucidating the full reaction co-ordinate, if the simulation can be done.

Lattice models have been employed by a number of workers. Gō *et al.* (1980), (see also Ueda *et al.*, 1978) have employed a series of simplified two and three-dimensional lattice models in which, for most cases, tertiary interactions are allowed only between residues in contact in the native state. They have successfully folded a three-dimensional lattice model of pancreatic trypsin inhibitor (PTI) from the denatured state, but were unable to do so for lysozyme. They ascribe this failure to the presence of mixtures of mirror image isomers. Recently, their work has focused on two-dimensional cubic lattice models (Taketomi *et al.*, 1988). They observed conformational transitions that are thermodynami-

cally and kinetically all or none. While their studies have pointed out the relative contribution of long and short-range interactions to the kinetics of folding, by allowing only native-like interactions between residues to occur, it is not clear how general their conclusions are.

More recently, Chan & Dill (1989) have exhaustively searched the sequence and conformational space of compact two-dimensional square lattice polymers and found that compact conformations are dominated by secondary structures. Furthermore, Krigbaum & Lin (1982) have used a bcc lattice model of PTI to investigate the relative folding efficiency of centrosymmetric *versus* a local interaction potential; both of which appear to work equally well in folding compact formations. Note that here too, a target potential is explicitly implemented in the folding algorithm.

Motivated both by the desire to deduce the general rules and by the necessity of reducing the number of degrees of freedom to make the folding problem tractable, we have adopted a minimalist approach to the protein folding problem (Kolinski *et al.*, 1986*a*,*b*, 1987*a*; Skolnick *et al.*, 1988, 1989*a*,*b*; Sikorski & Skolnick, 1989*a*,*b*, 1990) and in the initial stages of our studies, we have developed a series of diamond lattice models of globular proteins. Each α-carbon is represented by a bead on the diamond lattice and, for simplicity, two kinds of beads are considered. One kind is hydrophobic and the other is hydrophilic. Of course, beads representing all 20 different amino acids could be employed. This was not done, for two reasons. First, in the context of a minimal approach this was found not to be necessary. Second, one wants to keep the number of parameters to an absolute minimum, so as to make the results as convincing as possible. In addition to a hydrophobicity index, there is also a local, short-range marginal preference for β-states in β-proteins and for α-helical states in the folding of α-helical motifs. There are cases where the tertiary structure adopts a conformation different from the locally preferred one; see, for example, the requirement for formation of four-helix bundles with long loops (Sikorski & Skolnik, 1989*b*). Finally, there is a cooperativity parameter that mimics the effect of hydrogen bonding and local peptide dipole interactions. A key feature of the approach is that interactions between any pair of nearest-neighbor residues are allowed, and the native state is not specified in advance. To achieve the above objectives, we employ dynamic Monte Carlo (MC) algorithms (Skolnick *et al.*, 1989*a*) that must sort out the myriad of interactions and produce the same unique tertiary conformation on successive refolding. Otherwise, we would reject the current approach as non-viable.

To date, the equilibrium folding of a model, six-member Greek key analog of plastocyanin (Skolnick *et al.*, 1989*b*) and all variants of the left-handed, four-helix bundle motif (Sikorski & Skolnick, 1989*a*,*b*) have been successfully folded. Not only is a unique native conformation obtained, but the conformational transition is well approximated by a two-state model. We point out that an all-or-none transition is not an intrinsic property of the algorithm, but is the consequence of the interactions themselves (Skolnick *et al.*, 1989*a*).

At this juncture, a brief description of the dynamic MC method is appropriate (Binder, 1984, 1987). The system starts out in an arbitrary configuration, which is then subjected to successive local conformational rearrangements (micromodifications). Hence, MC sampling is employed to obtain the solution to a stochastic kinetics equation of motion that is conjectured to describe the dynamics of the system (Binder, 1987). This master equation of motion may be, for example, the Smoluchowski or Fokker Planck equation (Chandrasekhar, 1943). Provided that the sampling criterion satisfies detailed balance, in the limit that the number of micromodifications goes to infinity, the system will sample an equilibrium distribution of configurations. If one defines a "time" step as that when each individual piece of the system, on average, is subjected to all possible elementary micromodifications, then both thermodynamic and time-dependent averages can be obtained from a time average over the trajectory. The ability to obtain thermodynamic information follows from the ergodic hypothesis that ensemble averages may be replaced by time averages and *vice versa*. Observe that this method defines everything in terms of a reduced time-scale and, if only equilibrium sampling is of interest, it need not correspond to any physical time step. In our previous equilibrium sampling studies (Kolinski *et al.*, 1986*a*,*b*, 1987*a*; Skolnick *et al.*, 1988, 1989*a*,*b*; Sikorski & Skolnick, 1989*a*,*b*), the MC algorithm achieved its high efficiency by the simultaneous mixing of both long and short wavelength motions, thereby producing a distorted time-scale. As a consequence, the folding and unfolding pathways obtained from such an algorithm are suspect. Therefore, up to now, we have not reported any information on the mechanism by which the Greek key or the four-helix bundle motifs assembled. Recently, we have been able to surmount the distorted time-scale problem and have been able to fold our models employing elementary local moves alone. We report below the results from the folding unfolding pathways of model Greek key, β-barrel proteins. The accompanying paper reports similar results from the four-helix bundle motifs (Sikorski & Skolnicks, 1990).

The outline of the remainder of this paper is as follows. Section 2 presents a more detailed discussion of the model and the MC algorithm. The reader interested in qualitative insights alone can readily skip the latter subsection. Section 3 presents the simulation results for the folding and unfolding of the six-member, Greek key, β-barrel protein and then describes the free energy along the reaction coordinate constructed from a simple statistical mechanical theory. Section 4 summarizes the qualitative conclusions of this study and points out directions for future work.

2. Background

(a) *Model*

The model protein consists of a consecutive sequence of n α-carbon sites or beads on a tetrahedral lattice; each represents an amino acid residue that may be hydrophobic or hydrophilic in nature. By allowing no more than one residue to occupy a given lattice site, excluded volume is included. Furthermore, the choice of local moves (see below) is such that bond cutting does not occur, and therefore the effect of topological restrictions on folding is well accounted for. The conformational state of the model protein is given by a sequence of $n-3$ rotational states for the bonds, each of which may be in either the planar *trans* (t), or one of the two out of plane, *gauche* plus (g^+) or *gauche* minus (g^-) states. A t state corresponds to a β-state conformation, and a sequence of g^- states will produce a right-handed helix.

In order to proceed further, specification of the allowed interactions is required. We begin with the local, short-range interactions. Let ε_g be the intrinsic energy of a *gauche* state relative to a *trans* state. For amino acids involved in β-strand formation, it is not unreasonable to assume that ε_g is greater than zero. Since this parameter reflects short-range interactions, it is used as the basis of a reduced temperature scale, $T^* = k_B T/\varepsilon_g$, where k_B is Boltzmann's constant and T is absolute temperature.

Long-range interactions are assessed as follows. Imagine that a pair of residues (i and j) are non-bonded, nearest neighbors. As indicated in Figure 1, if both residues happen to be hydrophobic, then ε_h (a negative quantity) is the attractive potential of mean force that mimics (in a simple way) the hydrophobic interaction, taken in the quasichemical approximation. A potential of mean force is the effective interaction free energy between residues i and j when the solvent degrees of freedom are averaged over (Hill, 1956). Since ε_h is merely an effective attractive interaction parameter, it might also represent the reduction in free energy on formation of a salt bridge. Suppose, however, that one of the beads is hydrophobic and the other is hydrophilic. Then ε_w (positive) is the repulsive potential of mean force between them. Finally, we need to specify the interaction free energy when a pair of hydrophilic residues are non-bonded nearest neighbors. We have examined cases where the interaction is taken to be zero, slightly attractive or strongly repulsive (Skolnick *et al.*, 1989*a*; Sikorski & Skolnick, 1989*a*). Since the native state conformation in the models described below have all of their hydrophilic residues exposed, a potential that keeps them from associating is required. Thus, we take their interaction to be equal to ε_w as well; although qualitatively identical results have been obtained when their interaction is zero.

A final kind of interaction that has been introduced is a co-operativity parameter ε_c (Kolinski *et al.*, 1987*a*). This basically allows for non-bonded, second-nearest neighbor coupling when a pair of adjacent beads are associated with *trans* conformations. The explicit form of the conformational coupling is displayed in Figure 1. Thus, we allow both native and non-native interactions; that is, no target potential is introduced into the algorithm. By construction, there is a lowest free energy conformation; however, the algorithm has no *a priori* knowledge of this information. This is no less realistic than in the case of real globular proteins, where the chain must find the global free energy minimum by rattling about in configuration space, where all kinds of conformations are sampled. Note that in the primary sequences used below, based on short-range interactions alone there are 3^{12} isoenergetic turn conformations and a minimum of 16 isoenergetic loop conformations. Consistent with a given conformation, are a manifold of allowed tertiary contacts.

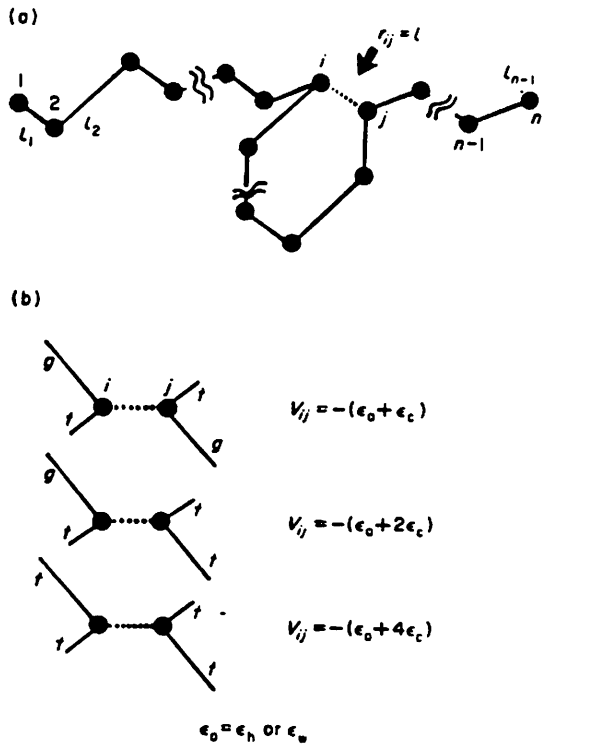


Figure 1. A representation of the allowed long-range interactions. Also displayed are the interactions embodied in the conformational co-operativity parameter ε_c.

The primary sequence is specified using the following convention. $B_i(k)$ is the ith stretch in the primary sequence that consists of k residues. The k residues have an identical ε_g; that is, they have a marginal intrinsic preference for *trans* ($\varepsilon_g > 0$) states. Possible bend regions are denoted by b_i and are located at the last two residues of stretch i and the first residue in stretch $i+1$. For convenience, $\varepsilon_b = \varepsilon_w = \varepsilon_c = 0$, but ε_g need not necessarily be zero. That is, putative bend regions are weakly hydrophilic. Putative loop regions are denoted by $L_i(k)$

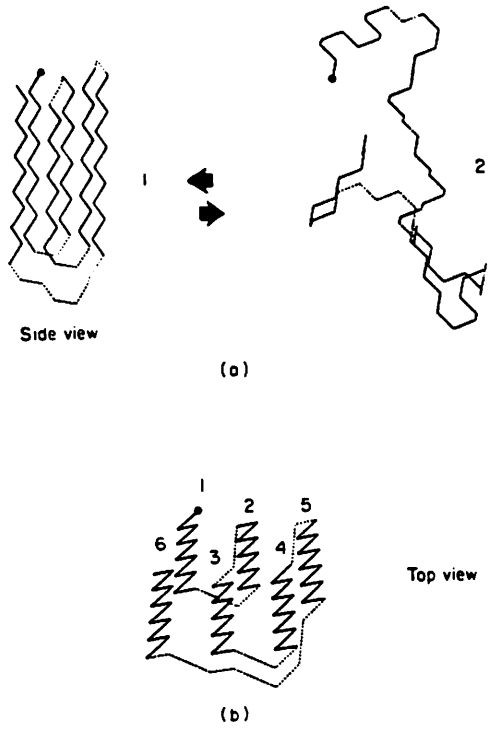


Figure 2. (a) Native state conformation of the 6-member, Greek key β-barrel 1 in equilibrium with a representative denatured state conformation 2. (b) Top view of Greek key β-barrel, showing the indexing of the β-strands. The circle denotes the N terminus.

and consist of k residues with $\varepsilon_c = 0$, but ε_h, ε_w and ε_g need not be zero.

The native state analog of a six-member Greek key β-barrel (1) is shown in equilibrium with a denatured state conformation (2) in Figure 2(a). Figure 2(b) shows a top view of the native conformation and the indexing scheme that labels β-strands 1 through 6. It should be pointed out that the potentials used above are all spherically symmetric. Thus, both the native conformation and its mirror image are isoenergetic; both have been obtained. The folding pathways are identical in character and thus, to avoid confusion, we present results for the physically correct conformation. Note that the introduction of side-chains breaks the mirror image symmetry. We have done this for a different lattice (unpublished results) and the problem of mirror image conformer degeneracy has been eliminated.

Elsewhere, we have shown that for an $n = 74$ model protein, a primary sequence pattern of the type

$$B_1(11)b_1B_2(11)b_2B_3(11)b_3B_4(12)b_4B_5(11)L(7)B_6(11)$$

produces an essentially all-or-none transition to the desired Greek key, 1 of Figure 2(a) (Skolnick *et al.*, 1989*b*). The inclusion of loop turn forming regions is consistent with the primary sequence of plastocyanin (Guss & Freeman, 1983). For B_i with $i = 1$ through 5, all the odd residues are hydrophobic and the even residues are hydrophilic; this is consistent with, but does not necessarily demand, β-sheet formation, as this juxtaposes hydrophobic residues as nearest neighbors in a β-hairpin conformation. However, due to the nature of a diamond lattice, for the native conformation, the even hydrophilic type residues located on strand 2 are nearest neighbors of the odd, hydrophobic type residues located on strand 5. A similar situation obtains for the contacts of strand 6, with strands 1 and 3, where the even hydrophobic residues in strand 6 are nearest neighbors to the even hydrophilic residues of strands 1 and 3. Thus, to allow for the possible stabilization of the native conformation due to hydrophobic interactions, the interactions described above between all these residues are set equal to ε_h. The aforementioned interactions hold not only for native-like conformations and contacts, but also for every conformation in which a given pair of residues are non-bonded, nearest neighbors.

The putative loop located at residues 57 through 63 is assigned a uniform attractive interaction of magnitude $\varepsilon_h = -\varepsilon_g$ with residues 33 and 34, which lie at the beginning of the turn between strands 3 and 4, in the native state. The loop is also assigned a repulsive ε_w with all the hydrophilic residues in strands 1, 2, 5 and 6. Furthermore, the local energetic preferences for residues 57 through 63 are $-2\varepsilon_g$, $2\varepsilon_g$, $-2\varepsilon_g$, $-2\varepsilon_g$, $2\varepsilon_g$, $2\varepsilon_g$, $-2\varepsilon_g$. A value $-2\varepsilon_g$ ($+2\varepsilon_g$) indicates that a g^+ or g^- (t) state is favored. Note that without long-range stabilization, the native loop conformation is one of 16 degenerate, lowest-energy states. Short-range interactions therefore do not enforce the native loop conformation. If the loop does not have at least a marginal preference for conformations (of the order of $\sim 2k_BT$) that include the native state, then the tremendous configurational entropy of the loop provides a manifold of out of register conformations of strand 6, namely in the transition region, it is not a faithful globular protein model (Skolnick *et al.*, 1989*b*).

In previous equilibrium studies, the simplest amino acid pattern that produced the unique Greek key 2 of Figure 2 and whose transition is all-or-none is model A having the primary sequence

$$B_1(11)b_1^0B_2(11)b_2^0B_3(11)b_3^0B_4(12)b_4^0B_5(11)L(7)B_6(11).$$

The hydrophilic residues in B_i had $\varepsilon_w = \varepsilon_g$, the hydrophobic residues had $\varepsilon_h = -\varepsilon_g/4$, and for all residues in B_i, $\varepsilon_c = -\varepsilon_g/2$. b_i^0 indicates that $\varepsilon_g = 0$ for those residues that might be associated with the turns; namely, residues 10 through 12, 21 through 23, 32 through 34 and 44 through 46. These are called "turn neutral", in that the native $g^+g^-g^+$ state is one of 27 isoenergetic states in the absence of long-range interactions. Thus, the model has no local turn bias for the native turn conformation whatsoever; however, it does not cost any free energy to form a turn in a b_i^0 region, whereas it does at other locations. This proves to be sufficent to localize the turns to these locations. The native

conformation has a configurational energy of $-70{\cdot}5\ \varepsilon_g$.

A natural extension of model A, which has turn neutral regions, is model B, where *trans* states are disfavored in putative turn regions. In model B, whose primary sequence is

$$B_1(11)b_1B_2(11)b_2B_3(11)b_3B_4(12)b_4B_5(11)L(7)B_6(11),$$

the turn neutral regions have been replaced by residues with an energetic preference for any *guache* state of magnitude $-2\varepsilon_g$. For the bends without long-range interactions, all of the eight triplets of *gauche* conformations are equally probable. Stabilization of the native $g^+g^-g^+$ comes from tertiary interactions. The native conformation for this model has a configurational energy of $-94{\cdot}5\ \varepsilon_g$.

Results are reported below for these representative choices of parameters. In fact, ε_c can be set equal to zero, if ε_h is augmented. The hydrophilic/hydrophilic contact parameter can also be set equal to zero, without qualitatively changing the results. As mentioned above, if the loop does not have any tertiary interactions in the native conformation, a number of out-of-register states of strand 6 is observed in the transition region. If the local stiffness parameter ε_g is increased too much, then artificially long single β-strands are observed in the denatured state (Kolinski *et al.*, 1986*a*,*b*). If $\varepsilon_g/\varepsilon_h \ll 1$ and $\varepsilon_w = 0$, then collapse to a non-unique random globule having patches of secondary structure occurs (Kolinski *et al.*, 1987*a*). In other words, we have defined the minimal requirements for the range of parameters, in a model protein, that produces a unique native state, obtained by an all-or-none transition from the denatured state.

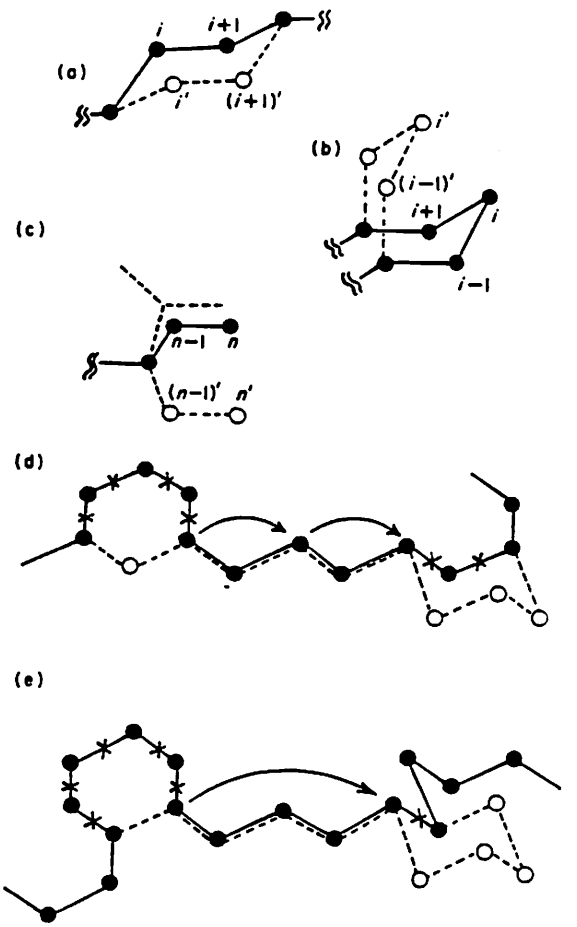


Figure 3. (a) Representative 3-bond kink motion. (b) Representative 4-bond kink motion. (c) Represenative 2-bond end motion. (d) Representative 4-bond wave motion previously employed for equilibrium sampling. (e) Representative 5-bond wave motion previously employed for equilibrium sampling.

(b) *Monte Carlo algorithm*

In what follows, we assume that the time-evolution of the model protein is described well by stochastic kinetics and employ a dynamic Monte Carlo technique to solve a master equation that has been shown to give the correct description of the dynamics of a random coil in the absence of hydrodynamic interactions (Kolinski *et al.*, 1987*b*). The use of a dynamic Monte Carlo method to provide qualitative insight into the dynamics of macromolecular systems has a long history in polymer physics (Baumgartner, 1984); the variant appropriate to the diamond lattice model is described in detail below. For the reader interested in the qualitative conclusions only, this section can be skipped.

The model is assumed to be subject to the following kinds of local rearrangements (Kremer *et al.*, 1981). (1) As depicted in Figures 3(a), there are three bond flip motions where the three bonds located in one half of the chair conformation of a cyclohexane-like ring jump to the other half; these serve to diffuse orientations down the chain. (2) Figure 3(b) displays four-bond kink motions involving the interchange $g^{\pm}\ g^{\mp} \rightarrow g^{\mp}\ g^{\pm}$. These serve to introduce new orientations into the chain. As pointed out by Boots & Deutch (1977), they are necessary, if excluded volume effects are included, to avoid an artificial and non-physical time-scale for the relaxation time of the end to end vector τ_R. (τ_R should scale as $n^{2\cdot2}$ rather than the non-physical result of τ_R going as n^3 if moves (2) are excluded.)

Moves (1) and (2) have been shown by Iwata & Kurata (1969) to span the space of all allowed motions on a diamond lattice comprised of three or more bonds. We also subject each of the ends to a two-bond end flip of the type depicted in Figure 3(c).

A single time step in the algorithm consists of the following.

(1) Store the energy of the old configuration, E_{old}.

(2) A bead is chosen at random and a three-bond flip is attempted.

(3) Then on a randomly chosen sequence of three beads, a four-bond flip is attempted. This process is repeated three times.

(4) Processes (2) and (3) are repeated $n-3$ times.

(5) Each of the ends is subject to a two-bond end flip.

(6) The energy difference ΔE between the old conformation, E_{old}, and the new conformation, E_{new}, is calculated. If $\Delta E < 0$, the new configuration is saved; if not, we apply the standard Metropolis criterion. That is, the probability $P = \exp(-\Delta E/k_B T)$ is calculated, and a random number, R, is generated. If $R < P$, the new configuration is saved; if not, the new configuration is rejected. This will generate a Boltzmann distribution of probabilities in the limit of a long sequence of steps.

Thus, per unit time, each bead is subjected, on average to a single, three-bond move and three four-bond moves; that is, the probability of a three-bond move P_{3b} equals 0·25 and the probability of a four-bond move P_{4b} equals 0·75. Thus, there are $3(1)+4(3) = 15$ attempts to move each bond, on average, per unit time. Elsewhere (Kolinski *et al.*, 1987*b*), we have shown that this choice of *a priori* probabilities generates correct Rouse-like dynamics with excluded volume (i.e. correct random coil dynamics in the absence of hydrodynamic interactions) for an athermal diamond lattice system. Moreover, it even generates the correct local dynamics are probed by standard orientational correlation functions (Valeur *et al.*, 1975). Thus, it appears that the essential features of local protein dynamics are preserved in these models; namely, local orientational diffusion, creation and annihilation, and therefore no gross distortion in the inherent time-scale is expected.

Let the mth bond connect beads $m-1$ and m; each vector on a diamond lattice has four bond vectors, either of the type **a**,**b**,**c**,**d**, or $-\mathbf{a},-\mathbf{b},-\mathbf{c},-\mathbf{d}$, with $\mathbf{a} = [+1,+1,+1]$, $\mathbf{b} = (1,-1,-1]$, $\mathbf{c} = [-1,1,-1]$, $\mathbf{d} = [-1,-1,+1]$. The stochastic dynamics process described above satisfies the following master equation for $P_a^m(t)$, the probability that the mth bond vector lies along **a** at a time t:

$$\frac{\partial P_a^m(t)}{\partial t} = W_3 \sum_{i,j,k,l \neq a} \{P_{ijiak}^{m-1} q_2(\mathbf{r}_{m-1}+\mathbf{a};\mathbf{r}_{m-1}+\mathbf{a}+\mathbf{i}) - P_{iaijk}^{m-1} q_2(\mathbf{r}_{m-1}+\mathbf{j};\mathbf{r}_{m-1}+\mathbf{j}+\mathbf{i}) + P_{kajji}^{m-3} q_2(\mathbf{r}_{m-3}+\mathbf{j};\mathbf{r}_{m-3}+\mathbf{j}+\mathbf{i}) - P_{kjiai}^{m-3} q_2(\mathbf{r}_{m-3}+\mathbf{a};\mathbf{r}_{m-3}+\mathbf{a}+\mathbf{i})\}$$
$$+ W_4 \sum_{\substack{i,j,k,l,p \neq a \\ p \neq i \\ p \neq j}} \{P_{ipijpk}^{m-1} q_3(\mathbf{r}_{m-1}+\mathbf{a};\mathbf{r}_{m-1}+\mathbf{a}+\mathbf{i};\mathbf{r}_{m-1}+\mathbf{a}+\mathbf{i}+\mathbf{j}) - P_{iaijak}^{m-1} q_3(\mathbf{r}_{m-1}+\mathbf{p};\mathbf{r}_{m-1}+\mathbf{p}+\mathbf{i};\mathbf{r}_{m-1}+\mathbf{p}+\mathbf{i}+\mathbf{j})$$
$$+ P_{kpjipi}^{m-3} q_3(\mathbf{r}_{m-4}+\mathbf{a};\mathbf{r}_{m-4}+\mathbf{a}+\mathbf{j};\mathbf{r}_{m-4}+\mathbf{a}+\mathbf{i}+\mathbf{j}) - P_{kajiai}^{m-3} q_3(\mathbf{r}_{m-4}+\mathbf{p};\mathbf{r}_{m-4}+\mathbf{p}+\mathbf{j};\mathbf{r}_{m-4}+\mathbf{p}+\mathbf{j}+\mathbf{i})\}, \quad (1)$$

wherein $P^m(t)_{i,j,k,l}$ is the probability that the mth bond lies along i, the $m+1$th bond vector lies along vector **j**, etc. and the sum is overall allowed orientations of the diamond lattice vectors. $q_2(\mathbf{r}_\alpha,\mathbf{r}_\beta)$ is the probability that the two contiguous lattice sites given by $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ are unoccupied. Similarly, $q_3(\mathbf{r}_\alpha,\mathbf{r}_\beta,\mathbf{r}_\gamma)$ is the probability that sites $\mathbf{r}_\alpha$, $\mathbf{r}_\beta$ and $\mathbf{r}_\gamma$ are unoccupied. If q_2 and q_3 are set equal to unity, the master equation first derived by Dubois-Violette *et al.* (1969) for three and four-bond motion on a diamond lattice in the absence of excluded volume is recovered. W_3 and W_4 are the *a priori* transition rates of three and four-bond motions, independent of the particular configuration of the chain. In the particular case chosen here, $W_3/W_4 = 1/3$. As explicitly shown in equation (1), excluded volume incorporates long-range interactions into the dynamics; that is, the jump cannot be made unless the sites into which the jump will be made are unoccupied. Thus, we have opted here for a numerical solution to the problem.

Replacing $\partial P_a^m/\partial t$ by $\Delta P_a^m/\Delta t$, and multiplying both sides of equation (1) by Δt, we see that $W_3 \Delta t$ is the fundamental unit of time for the simulation and corresponds to the transition rate of a three-bond jump in an athermal system.

Furthermore, the two end bonds satisfy:

$$\frac{\partial P_{ab'}^1}{\partial t} = W_2 \sum_{\substack{i \neq a \\ i \neq j}} \sum_{j \neq b'} \{P_{ij}^1 q_2(\mathbf{r}_3-\mathbf{b};\mathbf{r}_3-\mathbf{a}-\mathbf{b}') - P_{ab'}^1 q_2(\mathbf{r}_3-\mathbf{j};\mathbf{r}_3-\mathbf{j}-\mathbf{i})\} \quad (2a)$$

and

$$\frac{\partial P_{ab'}^{n-2}}{\partial t} = W_2 \sum_{\substack{i \neq a \\ i \neq j}} \sum_{j \neq b'} \{P_{ij}^{n-2} q_2(\mathbf{r}_{n-2}+\mathbf{a};\mathbf{r}_{n-2}+\mathbf{a}+\mathbf{b}') - P_{ab'}^{n-2} q_2(\mathbf{r}_{n-2}+\mathbf{i};\mathbf{r}_{n-2}+\mathbf{i}+\mathbf{j})\}. \quad (2b)$$

In practice, we have set $W_2 = W_3$.

A more detailed discussion of the master equation approach to stochastic dynamics has been presented by Binder (1987). Because master equations are constructed to satisfy detailed balance, they will generate an equilibrium Boltzmann distribution for the various states in the limit of a long Monte Carlo run. However, unless the elementary moves correspond to physical processes that occur on comparable time-scales, the dynamics may be patently non-physical. In previous work (Skolnick *et al.*, 1988, 1989*a*,*b*; Sikorski & Skolnick, 1989*a*,*b*), both four and five-bond wave motions (shown in Fig. 3(d) and (e), respectively), were permitted on the same time-scale as the local moves (Fig. 3(a), (b) and (c)); therefore, short and long wavelength motions were mixed and the time-scale was distorted. Here, only the smallest scale elemental steps, Figure 3(a), (b) and (c), consistent with the lattice description of the dynamics are used.

We next examine possible problems with the

dynamics that may affect the reliability of the qualitative features of the folding and unfolding pathways extracted from the trajectories. The most obvious defect in the algorithm is the apparent immobility of assembled secondary structures. For example, a linear β-sheet is immobile in the present algorithm, as there is no possibility of rigid body translations or rotations. However, in practice this restriction turns out not to be severe. The three-bond kink motions (Fig. 3(a)) can introduce a kink defect mechanism that allows pieces of *trans* stretches to translate. Similarly, an end rotation followed by a three-bond kink motion can effectively rotate *trans* stretches. While in practice this move is not particularly effective on isolated β-strands because such strands are marginally stable; β-hairpins have been seen to move about quite effectively by this mechanism. Moreover, the probability of a rigid body motion of a given piece of secondary structure should be exponentially damped relative to the elemental jumps by the relative number of residues involved in the motional units (Baumgartner, 1984). Small sections of isolated β-strands could move but are marginally stable; larger sections are more stable but also move far more slowly. Thus, this particular limitation is probably not important. Finally, due to limitations of computational resources, we have observed only a limited number of transitions; thus the N → D and D → N transition rates need not converge to their average values, and the reported transition rates are to be regarded as only approximate (but nevertheless reasonable).

3. Results

In the following, the nature of the folding and unfolding pathways in the model six-member, Greek key β-barrel is explored. The analysis proceeds from the examination of gross conformational properties to an examination of the finer details of representative trajectories that show the important events associated with folding and unfolding. Among the aspects explored in detail are the nature of the folding initiation events, the effect of local turn preferences on the folding and unfolding pathways, the nature of the transition state itself, and an approximate free energy analysis of the reaction co-ordinate.

(a) *Equilibrium averages*

Because of the local energetic preference for *gauche* states, the native conformation of model B has a lower free energy than model A. Thus, the N → D transition of model B should occur at a higher temperature than for model A. This is verified in Figure 4, where the mean-square radius of

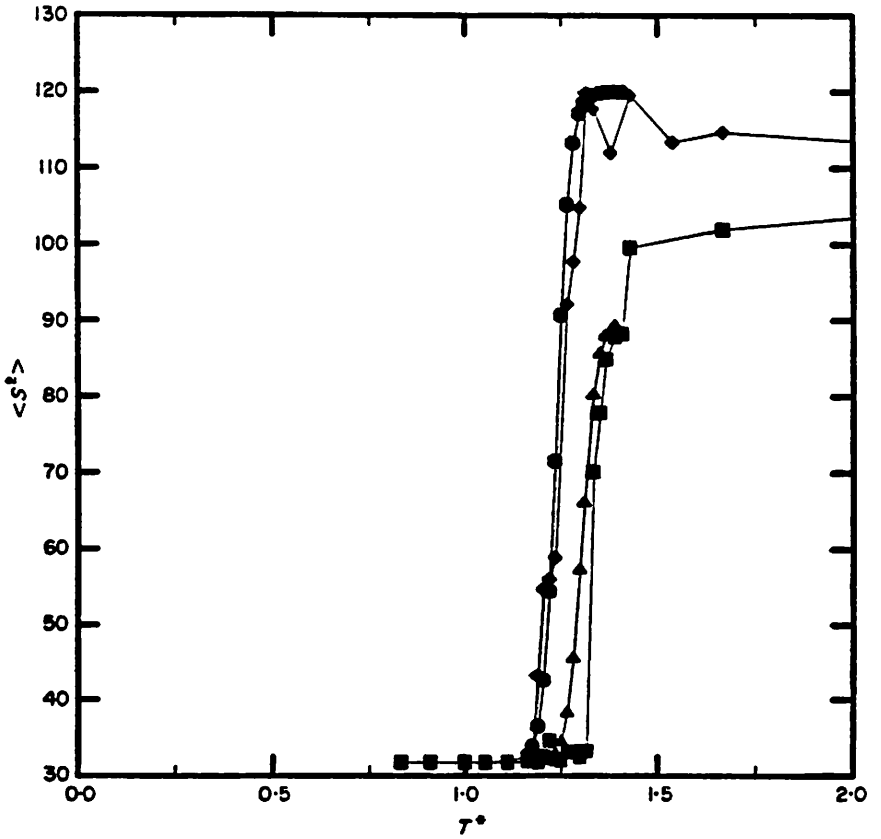


Figure 4. Plot of mean square radius of gyration $\langle \dot{S}^2 \rangle$ *versus* reduced temperature T^* for model A in the curves denoted by the (a) filled diamonds and (b) filled squares, respectively. (b) In the curves denoted by the filled circles (triangles) $\langle S^2 \rangle$ *versus* T^* is calculated *via* eqns (9) and (10) for model A.

gyration $\langle S^2 \rangle$ defined by:

$$\langle S^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} (r_i - r_{cm})^2, \quad (3)$$

with $|r_i - r_{cm}|$ the distance of the ith bead from the center of mass, r_{cm}, is plotted as a function of T^* for model A in the curve denoted by the filled diamonds and for model B in the curve denoted by the filled squares obtained from the equilibrium simulations.

For model A, the reduced transition temperature is at about $T^* = 1{\cdot}24$ and for model B it is at about $T^* = 1{\cdot}33$. The data of Figure 4 are obtained using the equilibrium sampling algorithm, which contains four and five-bond wave motions that distort the time-scale. While these moves increase the sampling efficiency by about a factor of 10 because both local and long wavelength moves are mixed, the folding pathways are suspect. (Actually, they turned out to be the same as when local moves alone were used.) Thus, we used these highly efficient moves to prepare an equilibrated configuration in the transition region. After this, the long wavelength moves are turned off, and only local dynamics are employed.

Tables 1 and 2 present for models A and B, respectively, configurational averages obtained from a series of runs in the order they were performed. Each run consists of 9×10^5 time steps. Column four summarizes the global conformational transitions (if any) associated with a given run; D is the denatured state and N is the native state. Columns five through seven give the average value of $\langle S^2 \rangle$, the fraction of *trans* states, f_t, and the average internal energy per residue, U. In columns eight through ten, f_{end}, f_{3b} and f_{4b} are the fraction of successful end flips, three-bond and four-bond motions, respectively. Observe that, although the probability of attempt of a three-bond move is only $\frac{1}{3}$ that of a four-bond move, the rate of acceptance of the three-bond moves is always higher. This reflects the difference in the intrinsic probability of finding a conformation suitable for three and four-bond moves, as well as the fact that three-bond moves require only two unoccupied sites to succeed,

Table 1

Summary of configurational properties for model A

Run no.	T^{*-1}	T^*	Conformation transitions	$\langle S^2 \rangle$	f_t	U‡	f_{end}	f_{3b}	f_{4b}
1	0·81	1·235	D → N	98·21	0·569	−0·0506	0·2451	0·11271	0·02181
2	0·82	1·220	N	31·82	0·765	−0·7569	0·03471	0·00153	0·00188
3	0·81	1·235	N → D	35·60	0·755	−0·7083	0·04909	0·00766	0·00298
4	0·81	1·235	D → N	79·45	0·617	−0·2247	0·19251	0·08503	0·01700
5	0·81	1·235	N	34·13	0·755	−0·7063	0·05798	0·00545	0·00323
6	0·81	1·235	N	32·77	0·760	−0·7247	0·04538	0·00370	0·00235
7	0·81	1·235	N	32·13	0·763	−0·7427	0·03880	0·00209	0·00201
8	0·81	1·235	N	34·52	0·754	−0·6995	0·06326	0·00647	0·00309
9	0·81	1·235	N	34·13	0·756	−0·7112	0·05838	0·00547	0·00269
10	0·81	1·235	N	32·94	0·761	−0·7279	0·04749	0·00365	0·00230
11	0·81	1·235	N → D	109·17	0·535	0·0785	0·28035	0·13381	0·02565
12	0·81	1·235	D	115·58	0·512	0·1403	0·29915	0·14424	0·02734
13	0·81	1·235	D	115·31	0·516	0·1405	0·30059	0·14399	0·02745
14	0·81	1·235	D → N	85·38	0·614	−0·2188	0·19436	0·08641	0·01699
Average of runs 1–14				62·22	0·674	−0·4223	0·1362	0·05302	0·01120
15	0·815	1·227	N	31·99	0·763	−0·7467	0·03818	0·00213	0·00197
16	0·795	1·258	N	34·51	0·752	−0·6871	0·06735	0·00651	0·00295
17	0·795	1·258	N	32·08	0·763	−0·7249	0·04348	0·00240	0·00209
18	0·79	1·266	N → D	103·99	0·547	0·0157	0·27108	0·12693	0·02428
19	0·795	1·258	D	121·55	0·516	0·1461	0·30810	0·14945	0·02796
20†	0·815	1·227	D	121·03	0·515	0·1425	0·30032	0·14511	0·02724
21	0·815	1·227	D → N	55·96	0·694	−0·4970	0·11196	0·04144	0·00911
22	0·79	1·266	N	32·66	0·759	−0·7098	0·04808	0·00305	0·00282
23	0·79	1·266	N	31·71	0·762	−0·7268	0·03908	0·00173	0·00196
24	0·815	1·227	N → D	35·59	0·751	−0·7059	0·05044	0·00862	0·00311
25	0·815	1·227	D	117·82	0·517	0·1404	0·29909	0·14399	0·02713
26	0·79	1·266	D → N	39·42	0·741	−0·6296	0·08631	0·01485	0·00460
27	0·79	1·266	N	35·63	0·747	−0·6583	0·08236	0·00914	0·00346
28	0·79	1·266	N	35·14	0·748	−0·6681	0·07720	0·00812	0·00328
29	0·785	1·274	N	34·82	0·751	−0·6724	0·07147	0·00717	0·00323
30	0·785	1·274	N	33·75	0·755	−0·6878	0·06375	0·00547	0·00274
31	0·785	1·274	N → D	66·59	0·655	−0·3497	0·16608	0·06276	0·01344
32	0·815	1·227	D	117·45	0·518	0·1273	0·29591	0·14033	0·02660
33	0·815	1·227	D → N	35·19	0·754	−0·7107	0·04965	0·00721	0·00294
34	0·785	1·274	N → D	39·31	0·738	−0·6299	0·08328	0·01479	0·00447

Each run consists of 900,000 time steps.

† The final configuration of run 18 served as the initial configuration of run 20.

‡ Average internal energy/residue.

Table 2
Summary of configurational properties for model B

Run no.	T^{*-1}	T^*	Conformational transitions	$\langle S^2 \rangle$	f_t	U‡	f_{1b}	f_{3b}	f_{4b}
1	0·75	1·33	N→D→N→D	50·48	0·682	−0·6854	0·13933	0·03739	0·00882
2	0·77	1·30	D→N	55·52	0·660	−0·6450	0·13848	0·04627	0·01050
3	0·75	1·33	N→D→N→D	46·17	0·693	−0·7251	0·13215	0·03088	0·00772
4	0·75	1·33	D→N→D→N	90·38	0·518	−0·2083	0·27703	0·11153	0·02251
5	0·75	1·33	N	32·67	0·757	−0·9123	0·06082	0·00397	0·00256
6	0·75	1·33	N→D	79·48	0·574	0·3691	0·22621	0·08767	0·01786
7	0·75	1·33	D→N	43·83	0·711	−0·7735	0·11488	0·02400	0·00628
8	0·75	1·33	N→D	82·92	0·546	−0·2908	0·25359	0·09797	0·01996
9	0·75	1·33	D→N→D	60·70	0·647	−0·5820	0·16776	0·05347	0·01181
Average†				60·82	0·641	−0·5683	0·17147	0·05586	0·01219

Each run consists of 900,000 time steps.
†Average over runs 1 and 3 through 9; i.e. at $T^* = 1{\cdot}33$.
‡ Average internal energy/residue.

whereas four-bond moves require three unoccupied sites. In all cases, f_{end} is greater than f_{3b} or f_{4b}, an entirely expected result. Observe further that all three of these quantities diminish substantially, by about a factor of 10 for f_{end}, and a factor of 30 for the f_{3b} and f_{4b}, in the native as compared to the denatured state. Note, however, that even in the native state there are conformational fluctuations mainly involving the ends.

(b) *Representative time-dependent statistics*

We next turn to representative time-dependent statistics. Figure 5(a) displays a plot of the instantaneous number of native contacts *versus* time obtained from runs 11 (top) and 14 (bottom) for model A; that is, a representative set of runs where successful transitions from the N to D and the D to N state are observed. The fully native molecule has 74 contacts. Fluctuations from this value involve the partial denaturation of strands 1 and/or 6. In Figure 5(b) and (c) we display, at a finer time resolution, the number of native contacts associated with the D → N transition of run 14 and the N → D transition of run 11. These are representative. Both reveal the presence of a marginally populated folding and unfolding intermediate having approximately 50 native contacts.

In Figure 6(a), we display a plot of the instantaneous value of the square radius of gyration S^2 *versus* time obtained from runs 11 and 14, and Figure 6(b) shows a finer time-resolution that focuses on the D → N transition region of run 14. In the denatured state, these are very large-scale fluctuations in S^2, characteristic of the broad sampling of configuration space. Observe that in the transition region there is a substantial diminution in the amplitude of S^2 prior to the formation of the native state. $\langle S^2 \rangle$ for the native conformation equals 31·7 in model units, where the distance between successive α-carbon atoms is $\sqrt{3}$. Figure 6(c) shows S^2 *versus* time in the N → D transition region for run 11. Similar behavior is observed for the average value of all configurational properties as a function of time. Based on these global properties, in particular the mean number of native contacts, it is clear for model A that the transition state involves an intermediate, I, having a substantial degree of native structure, but insufficient information has been presented to characterize the folding and unfolding intermediates.

In Figure 7, we plot the instantaneous number of native contacts N_c *versus* time obtained from run 3 of model B. Once again the presence of a native-like intermediate is revealed. This run also shows a number of unsuccessful unfolding attempts. Figure 8(a) shows S^2 *versus* time for run 3 of model B. Figure 8(b) and (c) show, on an expanded time-scale S^2 *versus* time associated with an N → I → N → D transition and a D → N transition, respectively. In Figure 8(b), the fluctuation in S^2 around $t =$ 125,000 arises from the random coil tails formed by strands 1 and 6 in the intermediate state, which then refolds around $t = 157{,}000$ to the native state. In all cases, plots of S^2 *versus* time in the transition region reveal a diminution in the amplitude of the conformational fluctuations prior to collapse in the native state. The nature of this transition state and the character of the folding initiation events need to be explored further.

(c) *Folding pathways*

Tables 3A and 4A present a compilation of the folding initiation events for models A and B, respectively, obtained by a detailed examination of the protein conformation as a function of time. Column two indicates that folding is seen to initiate at or very near to the turn between the pair of β-strands whose numbering is found in Figure 2(b). For example, for run 1 of model A, the initiation event occurred at the turn between β-strand 2 and β-strand 3; and is labeled 2,3 in column two. We define an initiation site as follows. The native conformation associated with the turn must occur at the start of a successful folding event from

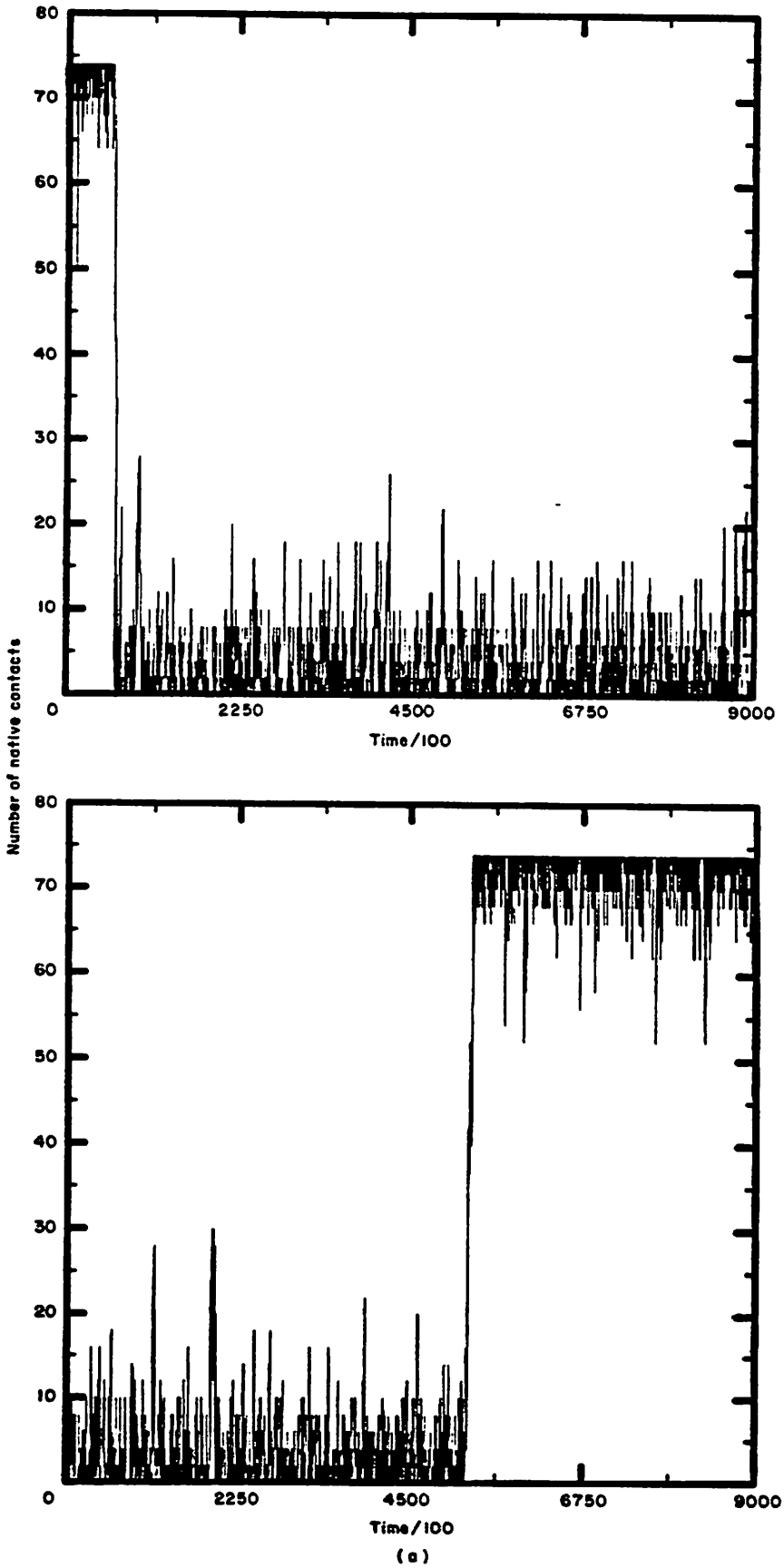

80
70
60
50
40
30
20
10
0
2250
4500
6750
9000
Time/100
Number of native contacts
80
70
60
50
40
30
20
10
0
2250
4500
6750
9000
Time/100
(a)

Fig. 5.

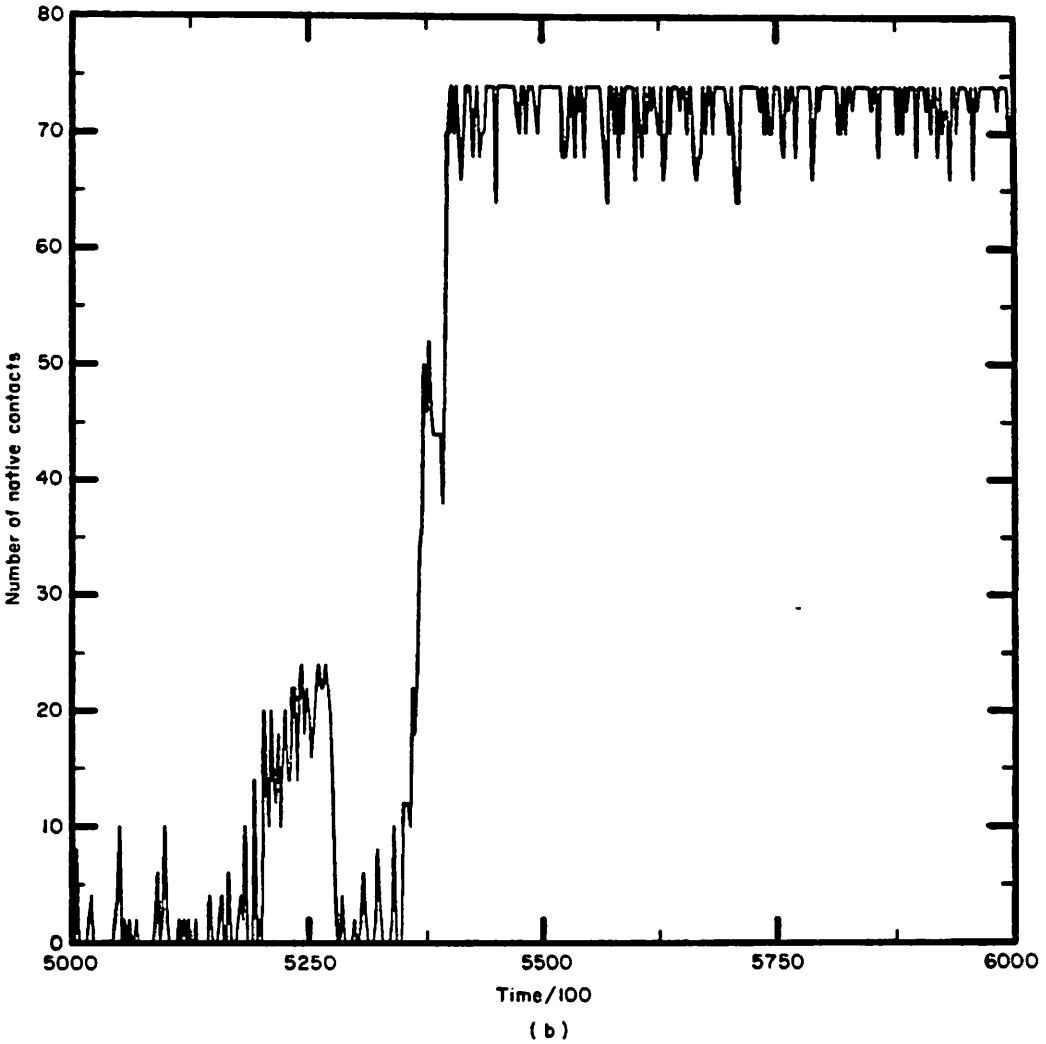


Fig. 5.

D → N, must be the first recognizable piece of secondary structure that exists in the native conformation, and must survive until the fully folded structure results. While more detailed snapshots consisting of a single time increment of representative trajectories have been examined frame by frame, for the purpose of this Table, we report results obtained from snapshots separated by 250 time steps. Since too few transitions have been observed to obtain more than qualitative insights, a course time-scale is appropriate. Better folding statistics are present in Table 4A than in Table 3A. Column three reports the time, τ_{DI}, from the initiation of folding until the folding intermediate consisting of the fully assembled, four-member β-barrel involving strands 2 to 5 is first obtained; that is, it is a *first passage time from the D → I state given that intermediate formation will occur*. Column four reports the first passage time, τ_{IN}, from I → N. Column five gives the ratio of τ_{DI}/τ_{IN}. In all cases, this ratio is less than unity, indicating that the rate-determining step is between I and N. This implies that a nucleation growth model (Wetlaufer, 1973) is inapplicable as a description of the pathway seen in these simulations. Once folding initiates it is not *downhill in free energy to the folded structure*. There are intermediates possessing substantial secondary structure having about 50 native contacts, before the native state is assembled.

For model A having neutral bends between all the β-stands joined by tight turns, out of a total of six folding initiation events, five successful folding events started at or near the turn between β-strands 2 and 3, and one initiation event was observed between strands 3 and 4. In the case of model B, Table 4A, where *gauche* states of any sign are preferred, of a total of eight successful folds, four involve initiation at the turn between β-strands 2 and 3, and three involved the turn between strands 3 and 4. Turn 2,3 involves residues 21 through 23. Turn 3,4 involves residues 32 through 34, and turn 4,5 involves residues 44 through 46. The reason that turn 2,3 occurs so often as an initiation site is twofold. First, in a number of cases, strand 1-2 first forms a hairpin that stabilizes β-strand 2, which then aids in initiation of folding from turn 2,3. The external strand 1 then rapidly forms and dissolves throughout the entire course of assembly (examples

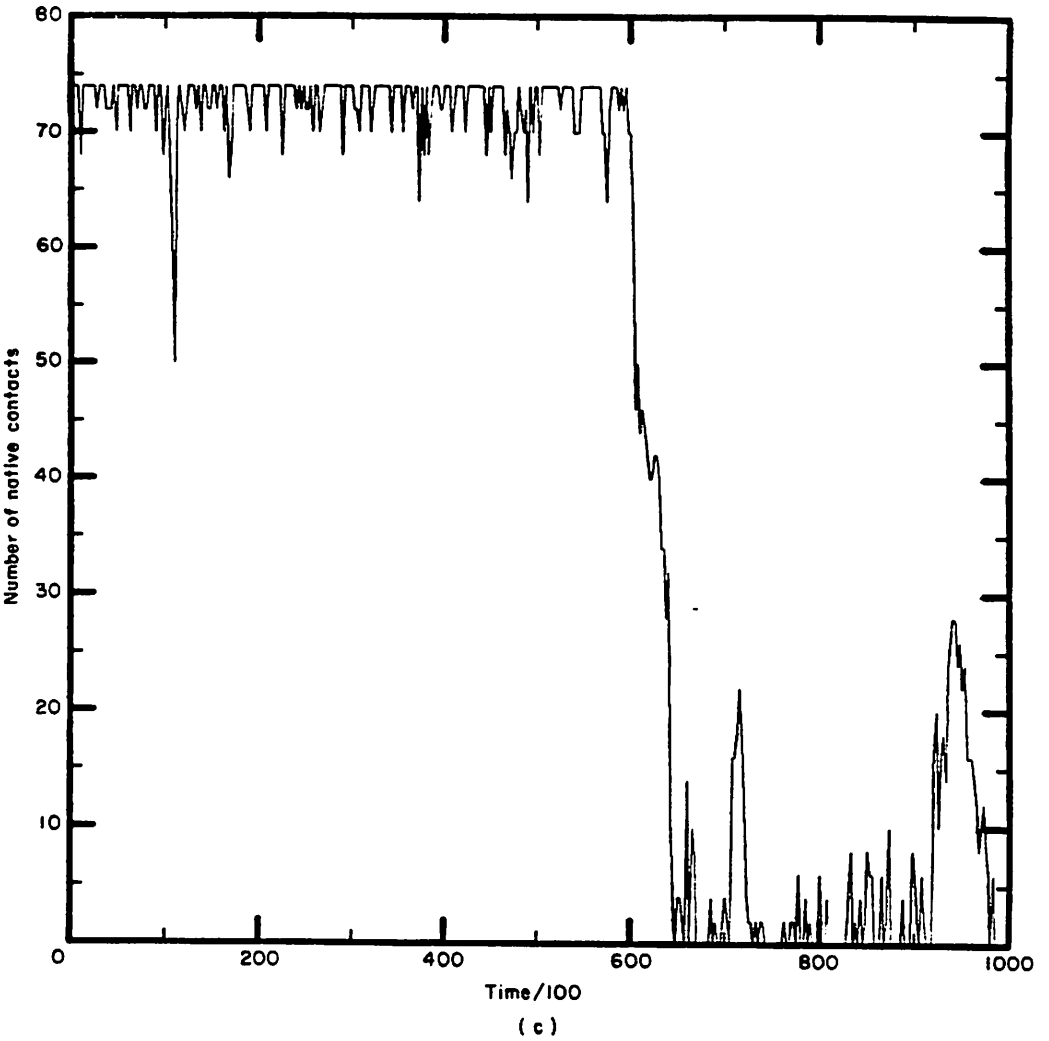
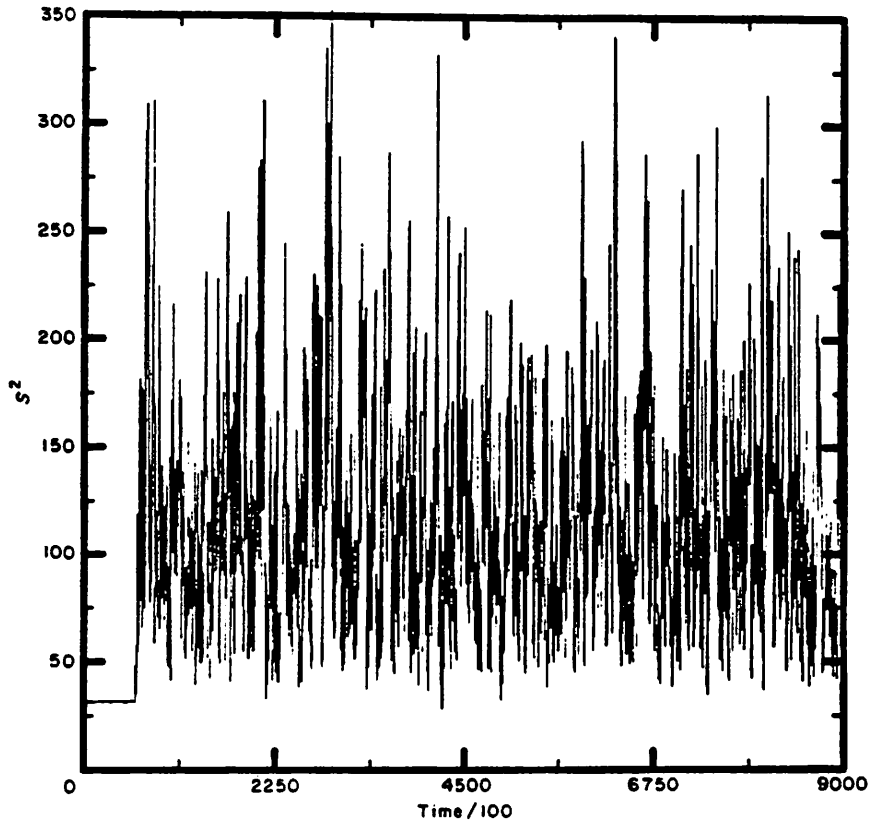


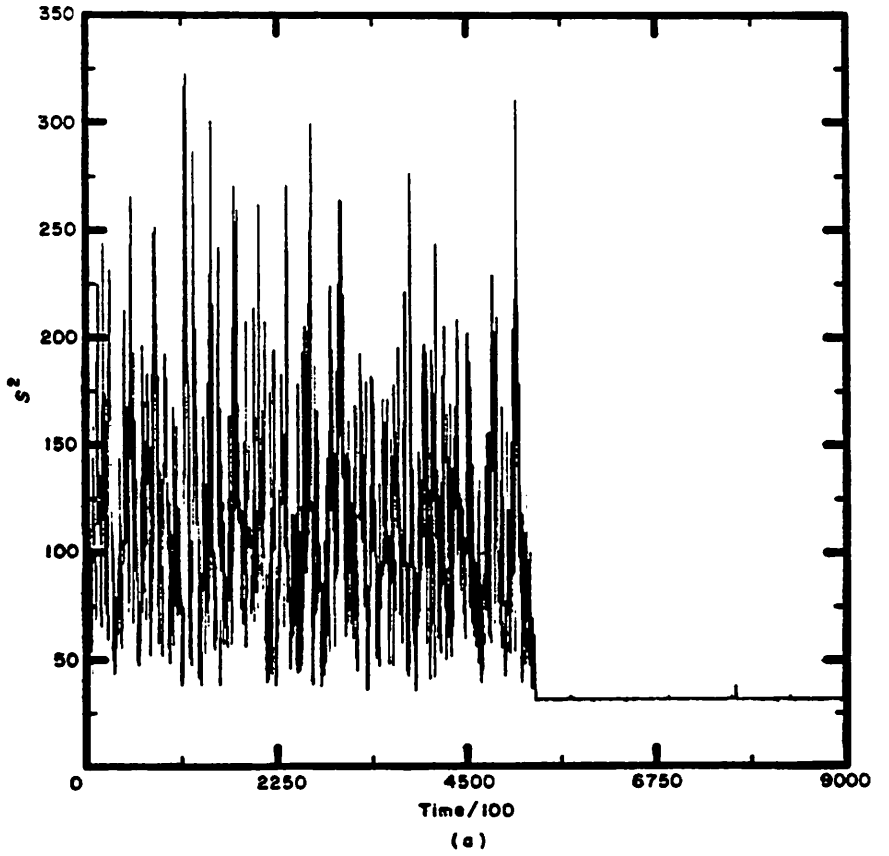
Figure 5. (a) Instantaneous number of native contacts *versus* time for runs 11 (top) and 14 (bottom) for model A. (b) The number of native contacts *versus* time at a finer time-resolution associated with the D → N transition of run 14. (c) The number of native contacts *versus* time at a finer time-resolution associated with the N → D transition of run 11.

of this behavior are seen below). Furthermore, due to the excluded volume effect (Teramoto *et al.*, 1958), turns are statistically more likely in a random coil at an end rather than in the middle. This rationalizes the observation that turn 2,3 is a somewhat more favorable initiation site, but what about turn 3,4? Actually, here the effect is much more subtle. For two of the three cases observed involving turn 3,4 initiation, folding actually began at strands 2-3. This structure then dissolved, leaving β-strand 3, which then rearranged to form 3,4. An example of this is seen in Figure 10.

Consistent with the previous equilibrium simulations results, the system spends a substantial fraction of its time either in the denatured state or in the fully native state, thereby making the transition thermodynamically all-or-none. For example, for model B at $T^* = 1{\cdot}33$, the total time spent unfolding, folding and in the intermediate state is 359,250 time steps out of a total of $7{\cdot}2 \times 10^6$ time steps or about 5% of the time.

In Figure 9, a representative folding trajectory extracted from run 14 of model A is shown. The time indicated in the Figure indicates the elapsed time from the start of the run. The reader is referred to $t = 540{,}250$ for the final assembled structure. The Greek key is shown from a side view and all snapshots are in the same perspective. At $t = 535{,}750$, strands 2 and 3 have successfully assembled. At $t = 536{,}000$, strands 2, 3 and 4 have adopted a native-like conformation. Observe, however, that β-strand 5 is in a non-native conformation, being colinear with strands 3 and 4, rather than colinear with strand 2. It takes until $t = 536{,}750$ before this incorrect fold has dissolved, and until $t = 537{,}250$ before the intermediate, now firmly identified as involving β-strands 2 through 5 has formed. During the course of assembly, β-strand 2 will dissolve and reform. See, for example, the snapshots at $t = 539{,}250$ and at $t = 539{,}500$. Thus, assembly is not unidirectional. At $t = 539{,}750$, the long loop plus strand 6 has almost worked its way into the


350
300
250
200
150
100
50
S^2
0
2250
4500
6750
9000
Time/100


350
300
250
200
150
100
50
S^2
0
2250
4500
6750
9000
Time/100

(a)

Fig. 6.

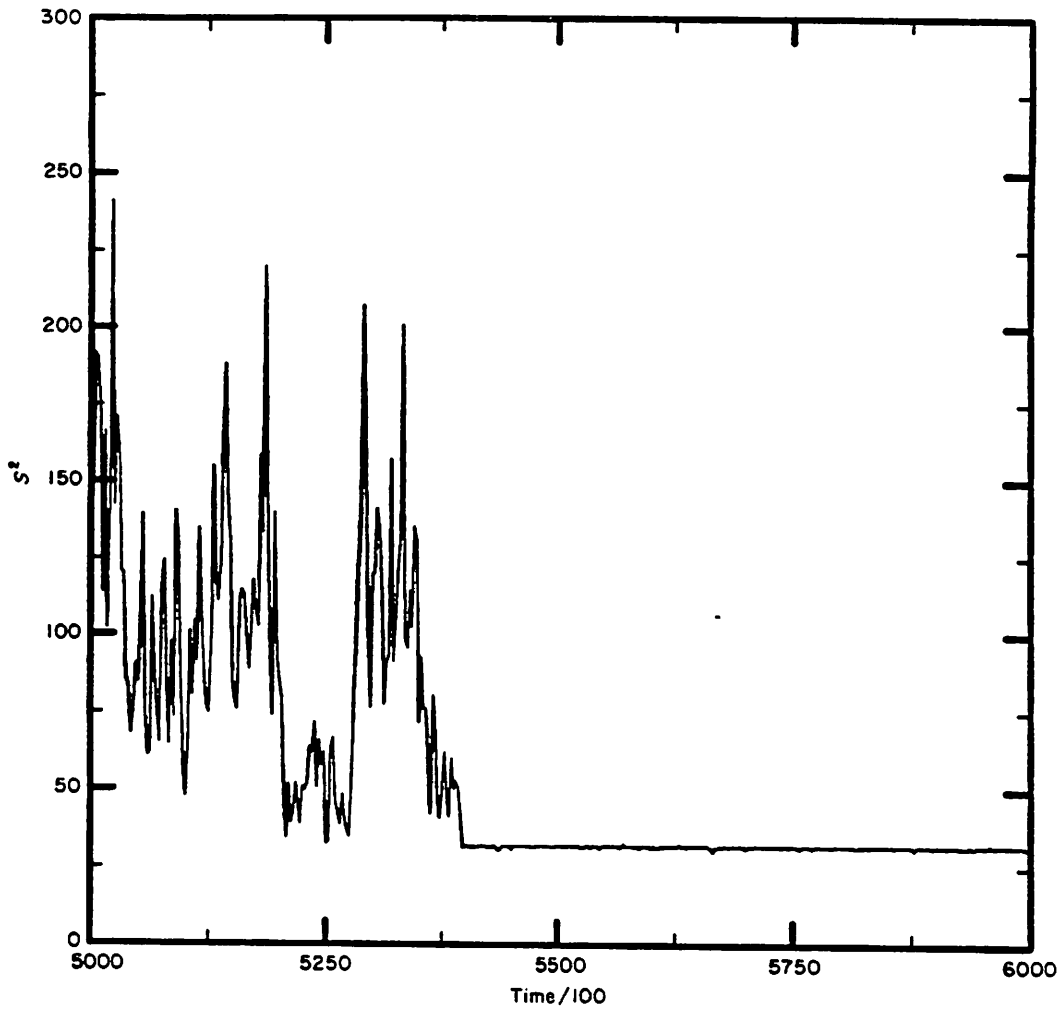


Fig. 6.

native state, and at $t = 540{,}250$, the lowest energy native conformation is obtained for the first time. Observe that the system spends the majority of its time in the intermediate state (3000 time units) as compared to the elapsed time for initiation of folding until the intermediate state first forms (1500 time units).

Figure 10 displays a representative folding trajectory obtained from run 4 of model B that shows the turn between β-strands 3 and 4 as the initiating site. At $t = 675{,}500$, a native-like turn between strands 2 and 3 has formed. By $t = 675{,}750$, β-strands 3 and 4 have fully assembled, and the turn between strands 2 and 3 is non-native; a partially formed β-strand 2 is coplanar with strands 3 and 4. Between $t = 675{,}750$ and 676,500, strands 3 and 4 partially dissolve and then reform to give the conformation shown at $t = 676{,}500$. By $t = 677{,}000$, strands 3 and 4 and the turn between strands 2 and 3 have reformed. By $t = 677{,}250$ strands 2 and 3 are native-like and strand 4 has partially dissolved. Finally, by $t = 677{,}500$, almost all of the residues in β-strands 2, 3 and 4 have adopted a native-like conformation. At $t = 677{,}750$ strands 2, 3 and 4 are almost native; the turn between strand 4 and 5 is non-native and strand 5 is clearly out of register. By $t = 681{,}500$, the four-member β-barrel folding intermediate is now intact. The time (i.e. at $t = 676{,}500$) from the formation of the first pieces of secondary structure that persist to the final folded state, until the formation of the intermediate is 5000 time steps. The long loop plus strand 6 will continue to thrash about until $t = 694{,}750$, an additional 13,250 time units, before the final native state assembly occurs.

The basic physical picture of β-barrel Greek key assembly that emerges from a detailed analysis of the above trajectories, as well as all the other trajectories compiled in Tables 3A and 4A is as follows. Folding tends to initiate at or near one of the β-turns; this is followed by the rapid assembly of a β-hairpin. The system assembles the secondary structure on site, with already formed secondary structure acting as scaffolding for subsequent native state structure formation. Thus, the β-strands zip up in place starting from the β-turn until the folding intermediate, the four member β-barrel is formed. Thus, the excluded volume effect exerted by the already assembled structure aids in subsequent

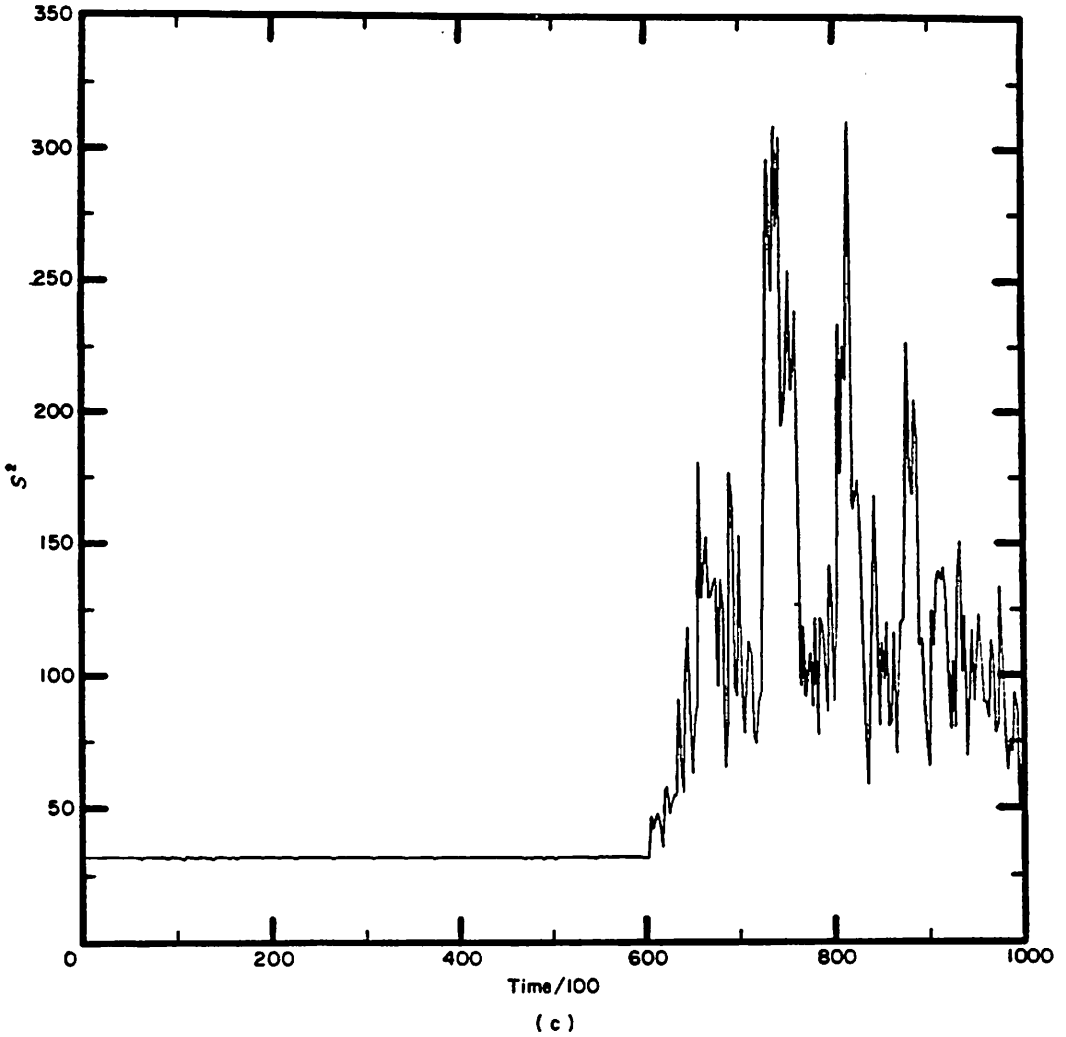


Figure 6. (a) Plot of the square of the instantaneous radius of gyration, S^2, *versus* time obtained from runs 11 (top) and 14 (bottom) for model A every 250 time steps over the entire run. (b) Finer time-resolution of S^2 *versus* time for the D → N transition of run 14. (c) Finer time-resolution for the N → D transition of S^2 *versus* time of run 11.

assembly. Once the intermediate forms, then the assembly is punctuated; the remaining long loop thrashes about trying to locate the very narrow pass in configuration space to the native state. (As we show below in a thermodynamic analysis of the reaction co-ordinate, the barrier on the intermediate side of the reaction co-ordinate is predominantly entropic in origin.) For example, the probability of finding the loop in the native conformation is 0·0135 without native contact stabilization, and with native contact stabilization it is 0·0607. Thus, even neglecting the excluded volume effect exerted by the assembled sections of the protein (see below), which will further slow down assembly, there is a large entropic barrier between the native and the intermediate state.

It should be further pointed out that, while sections of the protein are assembling secondary structure, especially in the early stages, the other portions of the native state may be dissolving. Compare, for example, in Figure 10 the conformation at $t = 677{,}000$ with that at $t = 677{,}250$. Thus, *assembly is not an irreversible process*. However, while there are a number of pathways to the intermediate, for example initiation at the turn between strands 2 and 3, 3 and 4, or 4 and 5 for that matter, the general process of assembly has increasingly limited possibilities the closer one is to the native state. This is not to say that folding from the intermediate state, I, to N is by a single unique spatial path, for it is not. For example, we have seen cases where the long loop at the end of strand 5 starts out on the side of strands 1 and 2, then eventually makes its way under the bottom of the folded structure before assembling. Here, the excluded volume exerted by the already assembled protein can drastically slow down folding. Other times, it is never on the wrong side of the barrel; rather, it works its way over strands 3 and 4 before assembly, etc. Thus, on the microscopic scale, many ways exist to assemble the protein, but in terms of a coarse description, a general pathway exists.

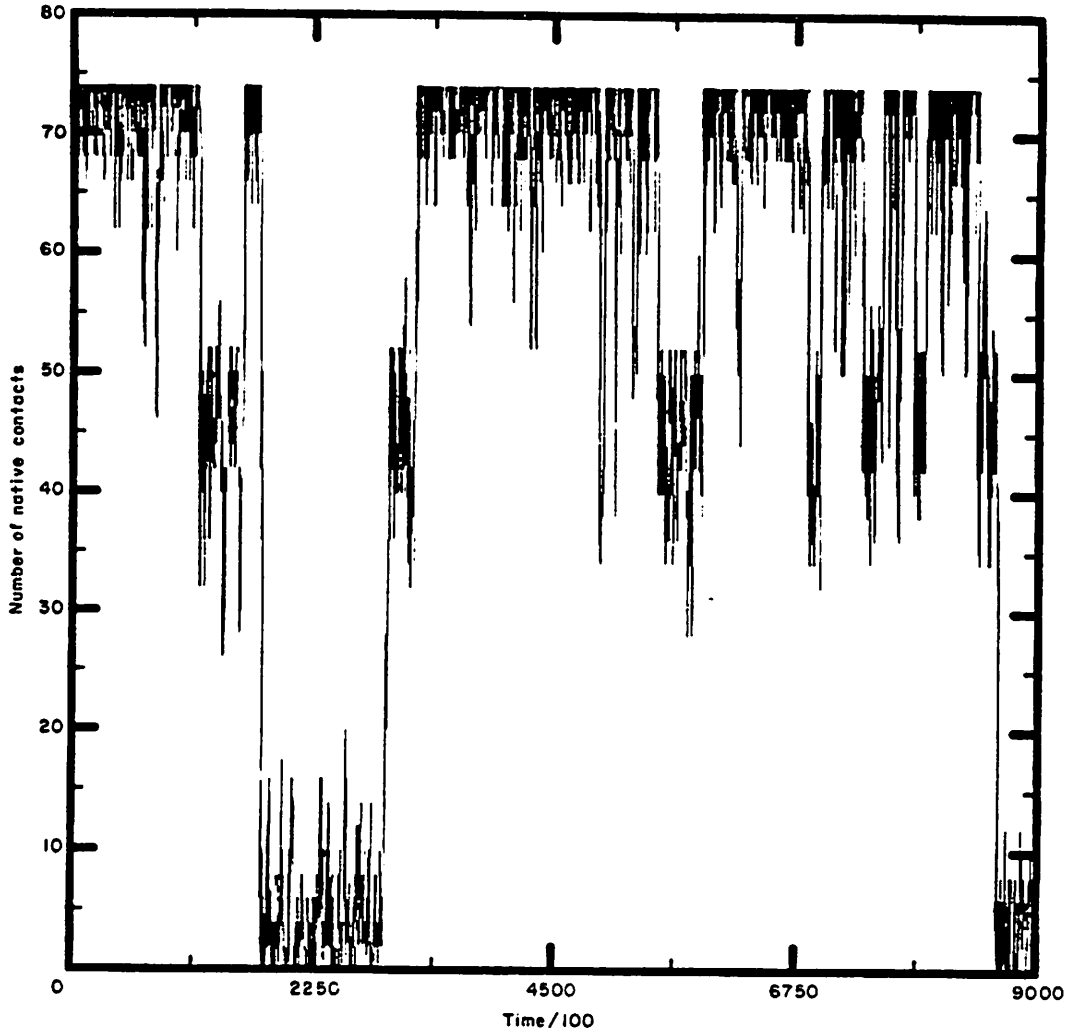


Figure 7. Instantaneous number of native contacts *versus* time obtained from run 3 of model B.

(d) *Unfolding pathways*

A compilation of the unfolding statistics is presented in Tables 3B and 4B for models A and B. Column two indicates the final element of tertiary structure of the native state that remains prior to dissolution to a fully random coil state. For example, 2,3 indicates that the turn between β-strands 2 and 3 is the last remnant of the native state that dissolves. τ_{NI} of column three is the approximate first passage time starting from a successful dissolution of strand 6 in the N state to the formation of the four-member β-barrel intermediate consisting of β-strands 2 through 5, with its attached random coil tails. τ_{ID} of column four is the approximate time required from the first appearance of the intermediate until no remaining tertiary structure remains. For model A, half of the final unfolding events involved the turn between strands 4 and 5 and the other half involved strands 3 and 4. For model B, of the nine unfolding transitions that have been observed, three out of nine involved final dissolution at the turn between strands 4 and 5, five out of nine involved strands 2 and 3, and one involved strands 3 and 4. The run in Table 3B at $T^* = 1{\cdot}227$, having $\tau_{NI}/\tau_{ID} = 4{\cdot}0$ was extremely atypical and involved the unraveling from end 1 due to the formation of a stable, colinear array of strands 2, 3 and 6.

In Figure 11, a representative unfolding trajectory is shown for run 18 of model A. Starting from the initiation of unfolding at $t = 132{,}000$, the intermediate state is reached by $t = 132{,}750$. The system then continues to thrash about until $t = 145{,}000$, after which rapid dissolution occurs. The final remaining piece of tertiary structure (which is shown at $t = 145{,}500$) is the bend between strands 4 and 5.

In Figure 12, we present a representative unfolding trajectory from run 34 for model A. Strand 6 in this case is seen to rapidly unfold from the native state (at $t = 841{,}500$) to the intermediate at 841,750; that is, in 250 time steps. The dissolved tail then thrashes about until $t = 850{,}250$, where strand 5 has dissolved and reformed in a non-native conformation. Dissolution of strand 2 is seen to

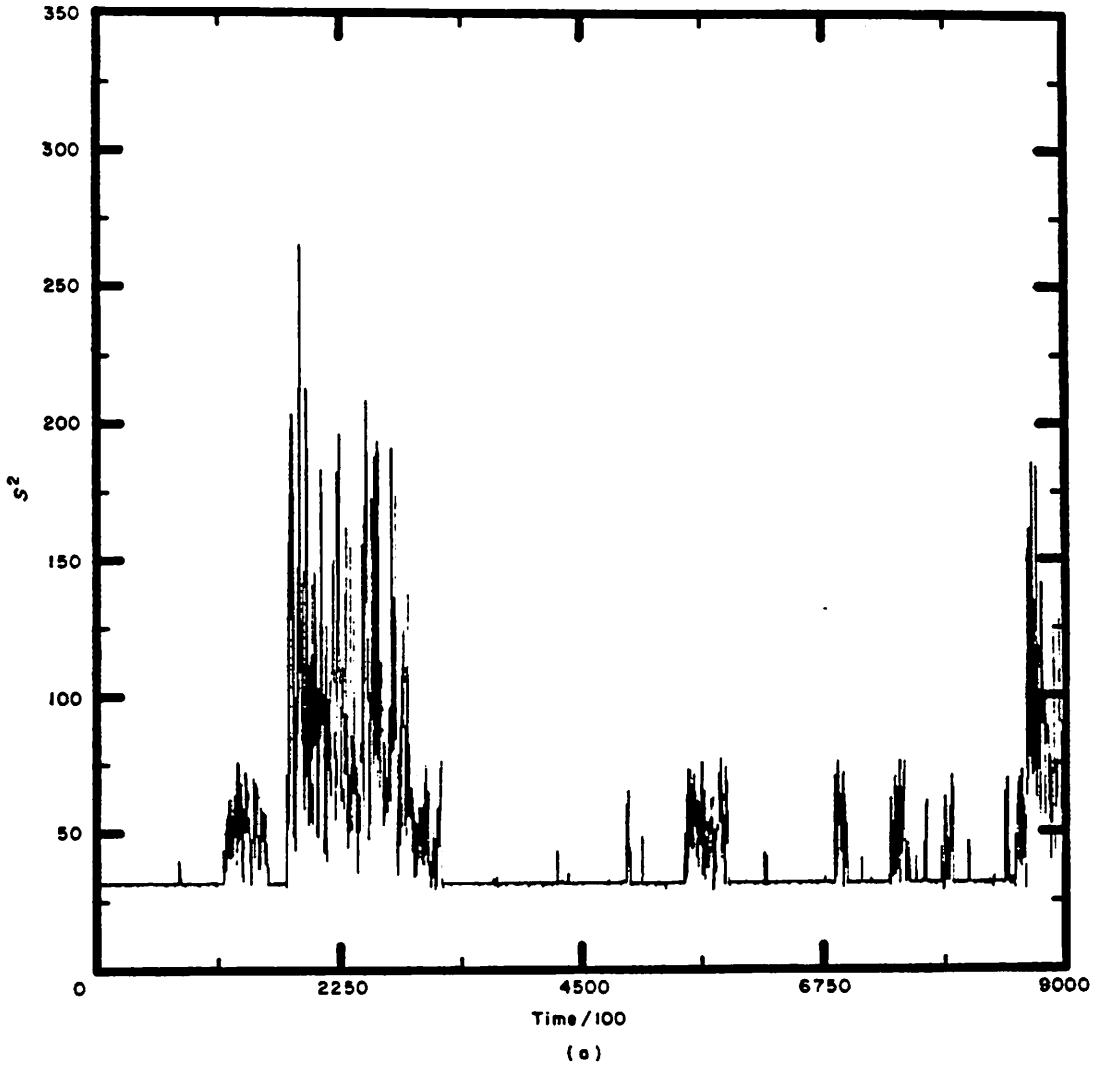


Fig. 8.

occur between $t = 851{,}000$ and $t = 851{,}500$. Finally, by $t = 851{,}750$, strands 3 and 4 have dissolved with no residual tertiary structure remaining after $t = 852{,}000$.

Figure 13 shows a representative unfolding trajectory obtained from run 9 of model B. Starting from the fully native conformation at $t = 745{,}750$, the intermediate forms very rapidly at $t = 746{,}250$. Dissolution of the four-member β-barrel from both ends of the C-terminal side is evident by $t = 757{,}250$. By $t = 758{,}250$, only the β-hairpin involving strands 2 and 3 is evident. All traces of the native conformation disappear by $t = 760{,}000$.

Qualitatively, the picture that emerges is that unfolding proceeds basically along the reverse pathway of folding. The key event in the initiation of an unfolding sequence is the dissolving of strand 6 from the native state. This is followed by the relatively rapid formation of the four-member β-barrel intermediate state involving β-strands 2 through 5. Further dissolution may proceed from either of the two ends of the chain, with all three turns 2 and 3, 3 and 4, and 4 and 5 having been seen as the location of the last residual tertiary structure. After unfolding, the dimensions are seen to expand as the system unfolds and begins to sample the broad expanse of configuration space characteristic of the denatured state.

(e) *Free energy of folding and unfolding*

Ideally, to substantiate further the qualitative picture of punctuated assembly, we would like to be able to extract directly from the simulation the free energy as a function of the reaction co-ordinate. Unfortunately, the statistics of folding are too poor to obtain a reasonable estimate for the entropy in the transition region from the simulations themselves. Adequate techniques do exist for extracting the conformational entropy outside the transition region (Miyazawa & Jernigan, 1982; Meirovitch *et al.*, 1988: these will be employed elsewhere), but for the transition from $D \rightarrow I \rightarrow N$, the conformational entropy is unreliable. Thus, we present below an approximate analysis designed to calculate the free energy along the reaction co-ordinate suggested by

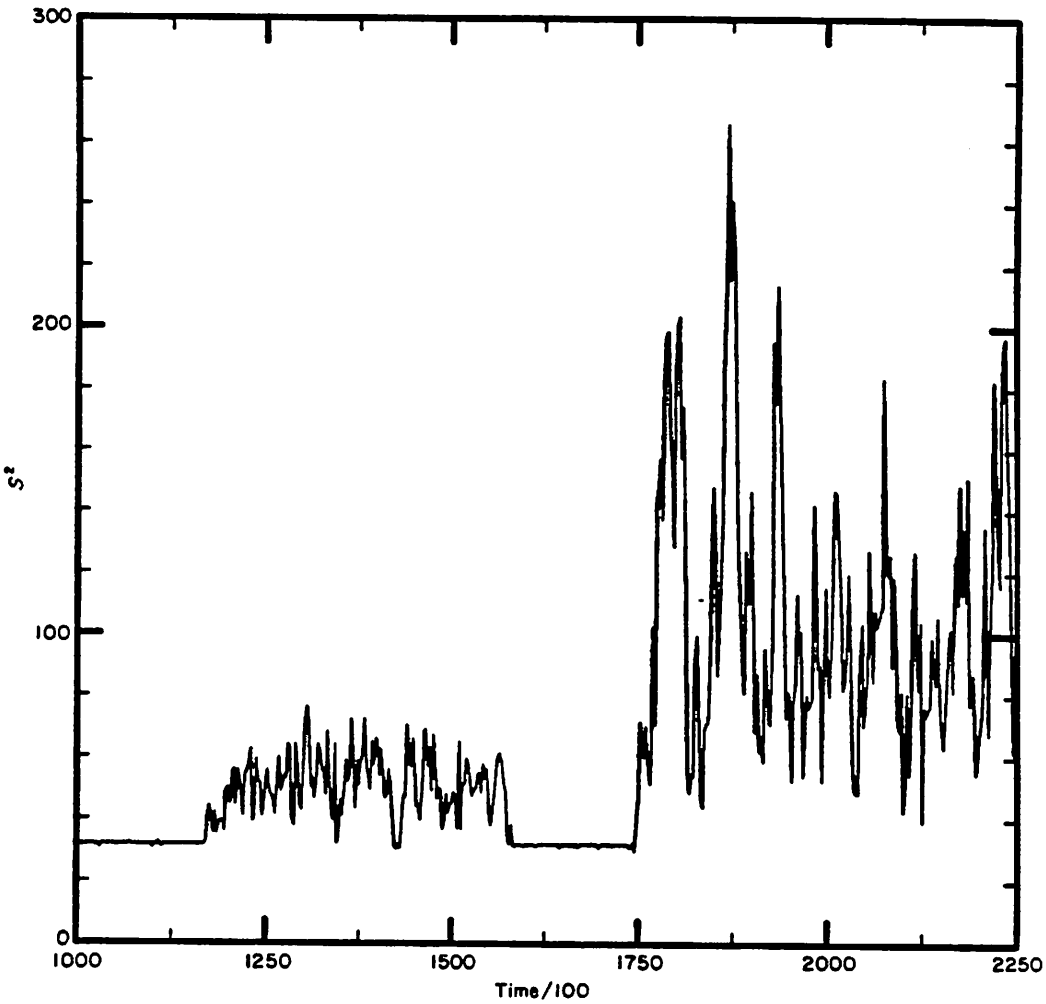


Fig. 8.

Figure 9; that is, β-strands 2 and 3 assemble first, this is followed by strands 3 and 4, and then the four-member β-barrel intermediate forms. Subsequently, strand 1 attaches to the native structure, followed by the loop adopting the native conformation and then strand 6 zips up into place.

The following assumptions are made. (1) The excluded volume effect is entirely neglected. (2) The denatured state is treated as a statistical random coil where all tertiary contacts are ignored. Thus, its internal configurational free energy (in units of k_BT) is:

$$A_D = -\sum_{i=1}^{n-3} \ln Z_i, \tag{4}$$

where Z_i is the rotational partition function of the ith bond:

$$Z_i = 1+2\,e^{-\varepsilon_i/T^*}, \tag{5}$$

ε_i is the rotational energy of the ith *gauche* state relative to ε_t. Thus, $\varepsilon_i = 1$, for example, for residues 2 through 9, 13 through 20, 24 through 31, 35 through 43, 47 through 56 and 64 through 72. For residues 57, 59, 60 and 63, $\varepsilon_i = -2$, and for residues 58, 61 and 62, $\varepsilon_i = +2$. For model B, (A) $\varepsilon_i = +2(0)$, for residues 10 through 12, 21 through 23, 32 through 34 and 44 through 46. For definity, we first focus on model B (qualitatively identical behavior follows for model A), whose random coil has an internal configurational free energy (in units of k_BT) of:

$$A_D = -(52 \ln (1+2\,e^{-1/T^*}) + 3 \ln (1+2\,e^{-2/T^*}) + 16 \ln (1+2\,e^{2/T^*})). \tag{6}$$

Thus an independent rotational isomeric states model for the calculation of the configurational free energy is employed, in which the zero of free energy is the all-*trans* state without any tertiary interactions (Flory, 1969).

(3) For a partially folded structure, we assume that the configurational free energy (in units of k_BT) of a conformation having N_C native contacts along the folding pathway is:

$$A_{N_C} = E_{N_C} - \sum_{i \neq N_c} \ln Z_i, \tag{7}$$

where E_{N_C} is the configurational energy (in units of

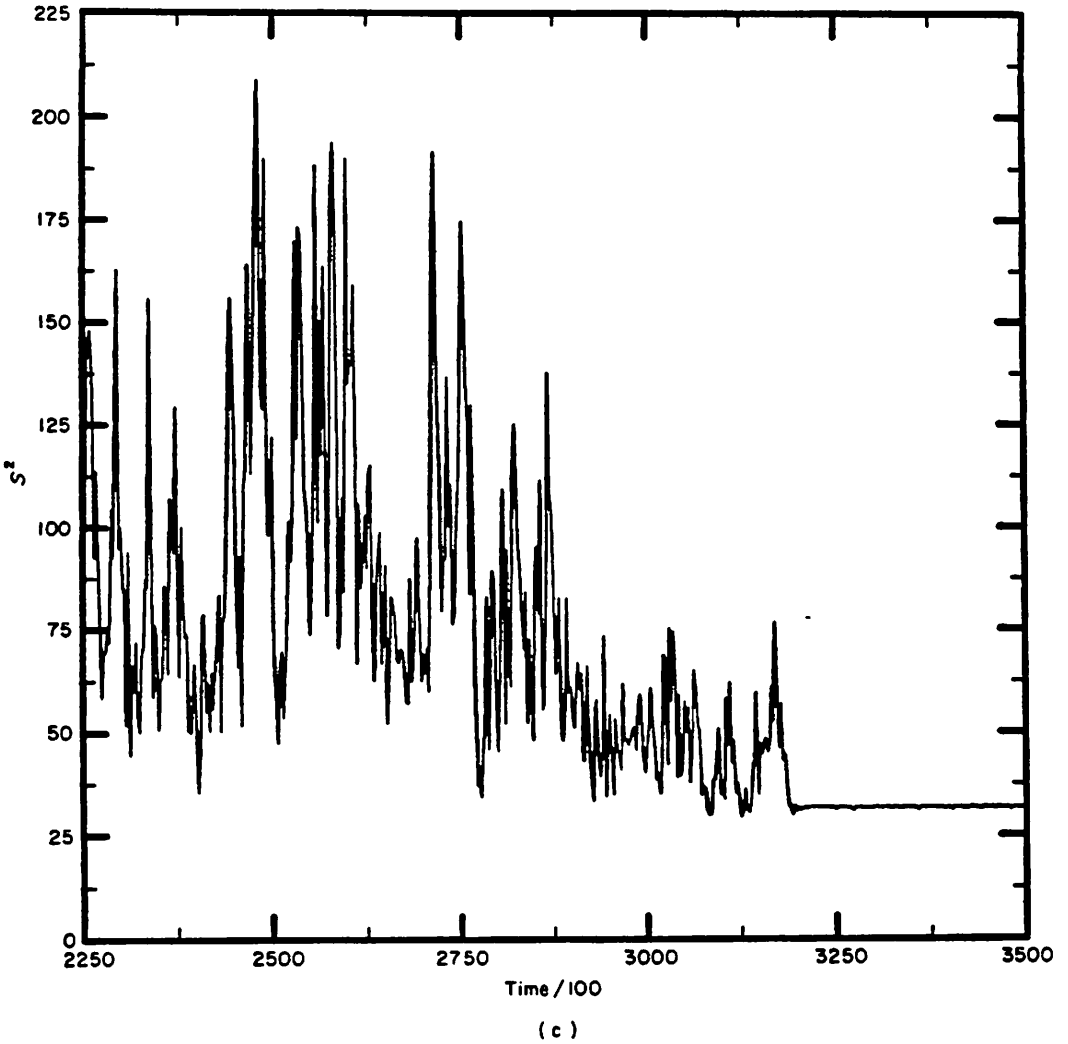


Figure 8. (a) Square of the instantaneous radius of gyration, S^2 *versus* time. (b) S^2 *versus* time on an expanded scale for N → D transition. (c) S^2 *versus* time on an expanded scale for D → N transition. In all cases, the trajectory of run 3, model B is employed.

k_BT) of the folded portion of the molecule having the specified N_c contact pairs, and the remainder of the unassembled chain is assumed to behave as if the partially folded structure is absent. The sum is over those portions of the molecule that have non-native conformations. See Table 5 for the conformations and energies associated with the various values of N_c. $N_c = 1$ corresponds to the formation of the native turn between strands 1 and 2, $N_c = 6$ corresponds to the formation of the turn between β strands 3 and 4. $N_c = 12$ corresponds to the formation of the native turn between strands 4 and 5. Since strand 5 has two neighboring β-strands (2 and 4), N_c increases by two each time β-strand 5 adopts two additional *trans* conformations. $N_c = 21$ corresponds to the additional formation of the native turn between strands 1 and 2, and $N_c = 26$ and 27 correspond to the formation of the native loop conformation. $N_c = 28$ and 29 correspond to the formation of the first pair of native contacts between β-strand 6 and strands 1 and 3. $N_c = 37$ is the fully native state.

Before employing the above approximations to calculate the configurational free energy along a reaction co-ordinate, one should be sure that, at the very least, the approximation is good for calculating the equilibrium between the native and denatured state. In Figure 4, the curves denoted by the circle and triangles give the results for the calculation of $\langle S^2 \rangle$ *versus* T^* for models A and B obtained *via*:

$$\langle S^2 \rangle = f_N \langle S_N^2 \rangle + (1 - f_N) \langle S_D^2 \rangle, \quad (9)$$

with N and D the denatured state values of $\langle S^2 \rangle$; f_N is obtained by:

$$f_N = \exp\{-[E_{37} - A_D]\}/\{1 + \exp(-(E_{37} - A_D)\}. \quad (10)$$

For model A, $\langle S_D^2 \rangle = 120$ and for model B $\langle S_D^2 \rangle = 90$. In both models $\langle S^2 \rangle_N = 31{\cdot}78$. Clearly, rather good agreement with the simulation is obtained.

In Figure 14(a), we plot $A_{N_c} - A_D$, the relative free energy of the partially folded molecule with respect to the denatured state *versus* the number of native contact pairs, N_C, at $T^* = 1{\cdot}54$, $1{\cdot}33$ and $1{\cdot}18$ for model B in the curves going from top to bottom;

Table 3
Compilation of statistics for model A

T^*	Folding initiation site	τ_{DI}†	τ_{IN}‡	τ_{DI}/τ_{IN}
A. *Folding statistics*				
1·235	2,3	750	2500	0·30
	2.3	1250	5000	0·25
	2.3	1500	4000	0·30
1·227	2.3	4250	18,250	0·23
	3,4	3250	17,000	0·19
1·266	2,3	2000	29,250	0·068
			26.000	
T^*	Final unfolding site	τ_{NI}§	τ_{ID}‖	τ_{NI}/τ_{ID}
B. *Unfolding statistics*				
1·23	4.5	250	5000	0·05
	4.5	500	4000	0·13
1·266	4,5	750	13,000	0·065
		500	—	
1·277	3.4	5000	1250	4·00
1·274	3.4	750	19,250	0·039
	3,4	250	10,250	0·024

† τ_{DI} is the time from successful initiation to the folding of the full 4-member β-barrel involving strands 2 through 5.

‡ τ_{IN} is the time from the 1st appearance of the fully assembled 4-member β-barrel involving strands 2 through 5 to the fully assembled native Greek key.

§ First passage time from the start of unfolding in the fully native conformation to the 4-member β-barrel intermediate involving strands 2 through 5.

‖ Time for complete dissolution of the 4-member β-barrel intermediate.

Table 4
Compilation of statistics for model B

T^*	Folding initiation site	τ_{DI}†	τ_{IN}‡	τ_{DI}/τ_{IN}
A. *Folding statistics*				
1·333	3,4,5§	750	34,000	0·022
			40,750	
	2,3	3000	26,000	0·12
			40,500	
			12,750	
	3,4	5750	13,250	0·434
	2,3	750	19,250	0·039
	3,4	3500	13,500	0·26
	2,3	250	5250	0·048
Average		2333	22,806	0·10
1·299	3.4	1000	15,000	0·067
1·176	2,3	7000	80,000	0·088
T^*	Final unfolding site	τ_{NI}‖	τ_{ID}¶	τ_{NI}/τ_{ID}
B. *Unfolding statistics*				
1·33	2,3	1000	11,750	0·085
	4,5	2000	13,250	0·15
	—	1250	—	—
	4,5	750	2000	0·375
	—	1250	—	—
	4,5	750	19,750	0·038
	—	500	—	—
	2,3	1000	25,250	0·040
	2,3	1250	15,000	0·083
	3,4	3500	42,250	0·0061
	2,3	500	13,750	0·036
Average	—	1250	17,875	0·059
1·205	2,3	3750	10,500	0·367

† τ_{DI} is the time from successful initiation to the folding of the fully assembled 4-member β-barrel involving strands 2 through 5.

‡ τ_{IN} is time from the 1st appearance of the fully assembled 4-member β-barrel involving strands 2 through 5 to the fully assembled native Greek key.

§ The initiation site in the run was ambiguous. The 1st identifiable structure for $\Delta t = 250$ was β-strands 3, 4 and 5.

‖ First passage time from the start of unfolding in the fully native conformation to the 4-member β-barrel intermediate involving strands 2 through 5.

¶ Time for complete dissolution of the 4-member β-barrel intermediate.

that is, under conditions where the denatured state is favored, in the transition region, and under strongly native conditions. The first maximum in the $A_{N_C}-A_D$ plot occurs after the β-hairpin involving strands 2 and 3 has formed, but before the β-strand 4 has fully assembled. At high values of T^*, the entropic term dominates and it is very easy for the β-hairpin involving strands 2 and 3 to dissolve; that is, it is on a free-energy ledge, not in a local minimum. Thus, under denaturing conditions, it is extremely difficult to stabilize early folding intermediates. As T^* decreases, the energetic term starts to dominate, and then the structure containing three β-strands becomes trapped in a local minimum. Observe that the lowest free-energy state is not the fully assembled triplet of β-strands 2, 3 and 4, but involves a conformation where the last two residues of β-strand 4 are frayed. At all temperatures, there is always a barrier of entropic origin between the three-stranded structure (2,3,4) and the four-member, β-barrel of strands 2, 3, 4 and 5 that occurs after the first native contact in β-strand 5. This is responsible for the punctuated assembly we have observed in preliminary studies of four-member, β-barrel folding. The barrier at $N_C = 12$ between the three-member β-barrel involving 2,3,4 (four-member, β-barrel state) and the four-member β-barrel state with $N_C = 18$ (3-member barrel) decreases (increases) from 2·98(3·39), 2·52(5·27), 2·07(7·19) as T^* decreases to 1·54, 1·33 and 1·18, respectively. This reflects the fact that the free-energy barrier from the D side is mainly entropic, while from the I side, the barrier is predominantly energetic. Thus, the four-member β-barrel intermediate becomes more accessible as T^* decreases, and it becomes increasingly difficult to dissolve the intermediate as T^* decreases.

Observe further that there is a broad free-energy valley evident between $N_C = 18$ and $N_C = 26$; this corresponds to the assembly of β-strand 1 and formation of the first native contact between the long loop and the previously assembled native state. The

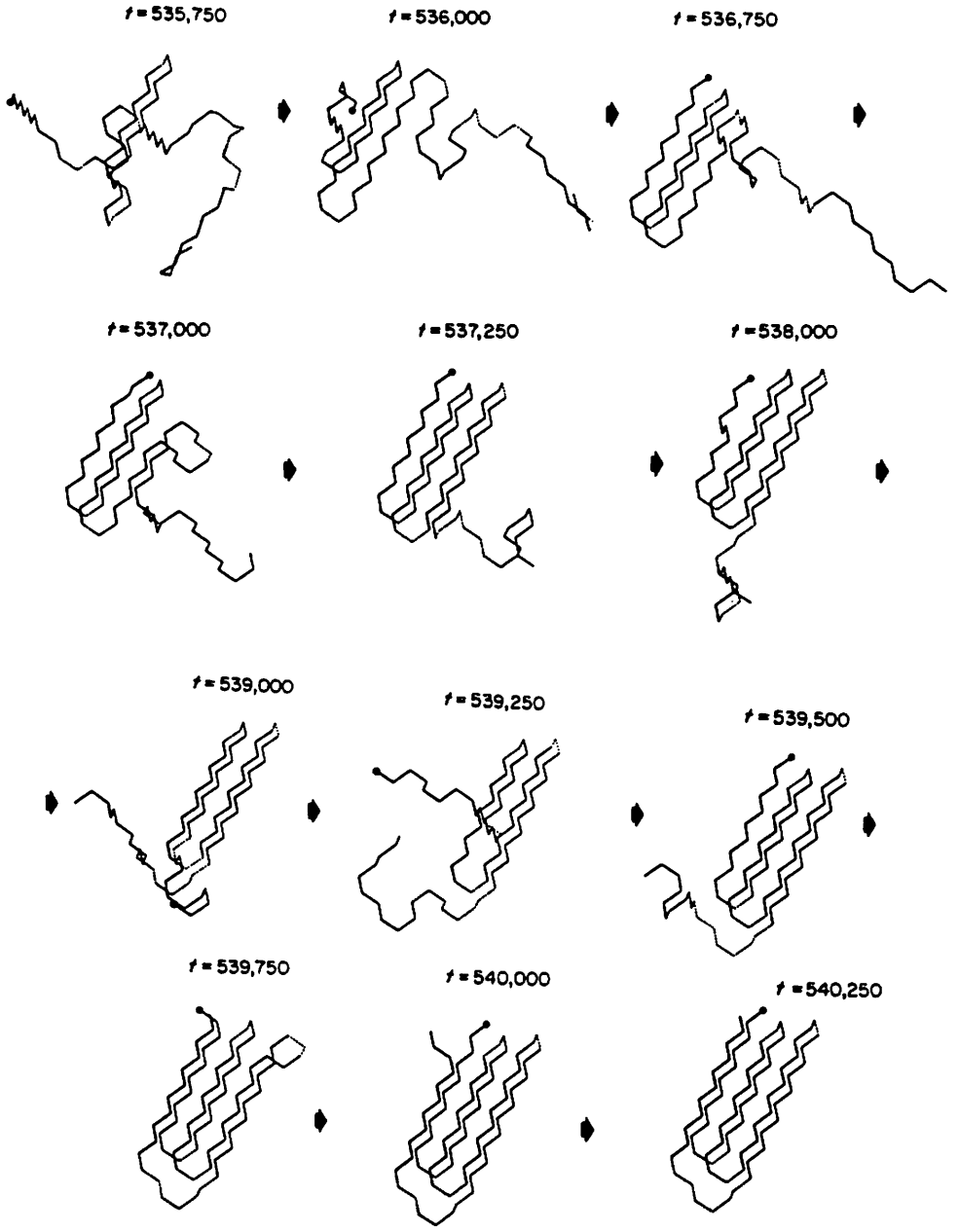


Figure 9. Representative folding trajectory extracted from run 14 of model A is shown. The circle denotes the N terminus.

existence of this broad valley, separated by a large entropic barrier from the fully native conformation is the origin of the long pause always observed in the simulations after the four-member β-barrel was assembled. The absolute height of this barrier, which is always located at $N_C = 29$, equals 15·74 at $T^* = 1{\cdot}54$, equals 8·19 at $T^* = 1{\cdot}33$ and equals 0·85 at $T^* = 1{\cdot}18$. Thus, it becomes increasingly easier to find the native conformation as the temperature is lowered. In other words, at higher temperatures where the configurational entropy dominates, it is very difficult to surmount the free-energy barrier from the intermediate to the native state; more explicitly, the barrier calculated as the difference in free energy between $N_C = 26$ and $N_C = 29$ equals 4·44 at $T^* = 1{\cdot}54$, 4·10 at $T^* = 1{\cdot}33$ and 3·78 at $T^* = 1{\cdot}18$. Thus, as the energetic stabilization of the native state dominates, the barrier due to entropic contributions between the intermediate and the native state decreases with decreasing temperature. In other words, folding from the intermediate to the N state becomes easier as the temperature decreases.

Conversely, the height of the barrier between the N state and the intermediate (which strictly speaking is seen to be the 4-member barrel plus a population of states including various degrees of assembly of the external strand 1, both with and without the first native loop contact) increases from 4·74 at $T^* = 1{\cdot}54$ to 6·60 at $T^* = 1{\cdot}33$ and to 8·43 at

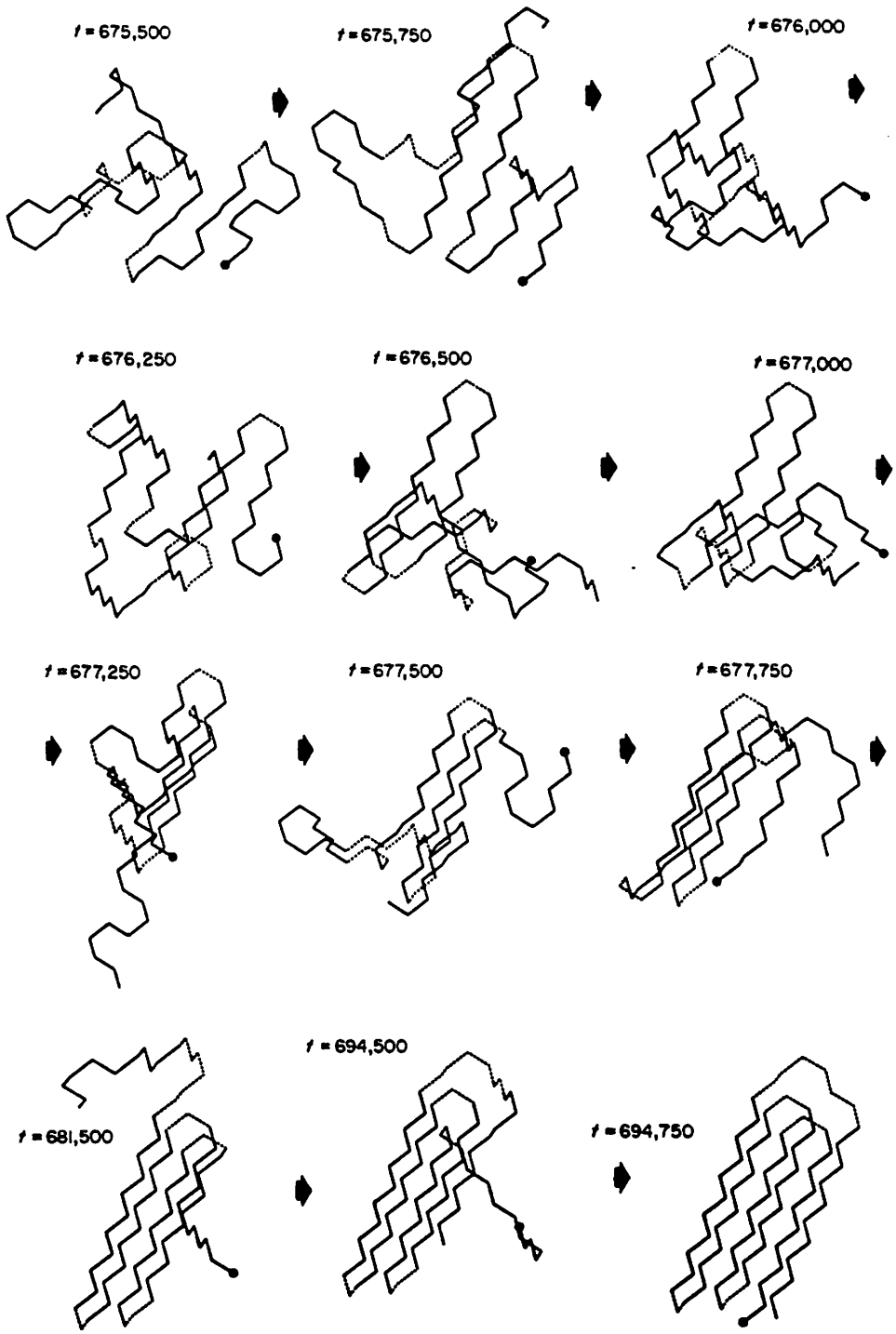


Figure 10. Representative folding trajectory extracted from run 4 of model B. The circle denotes the N terminus.

$T^* = 1{\cdot}18$. The relative free energy of the N state is 10·99 at $T^* = 1{\cdot}54$, 1·60 at $T^* = 1{\cdot}33$ and −7·58 at $T^* = 1{\cdot}18$. Thus, the conjecture of Goldenberg & Creighton (1985) that the native conformation is kinetically trapped is borne out in this model calculation. However, unlike their Cardboard Box model of protein folding, which conjectures that the transition state is a high-energy distorted form of the fully folded state, in this model, the transition state consists of 29/37 native contact pairs. The transition state is identified as the six-member β-barrel, all of whose turns and loops are native, and which has the first two residues of β-strand 6 closest to the N terminus in the native state; thus, the transition state lies very close to the native state as required by experiment. In Figure 14(b), we plot in the lower and upper curves (filled circles and squares) E_{N_C} and $A_{N_C} - E_{N_C} - A_D$ *versus* N_C at $T^* = 1{\cdot}33$. The latter term is the difference in configurational free energy associated with converting the random coil state

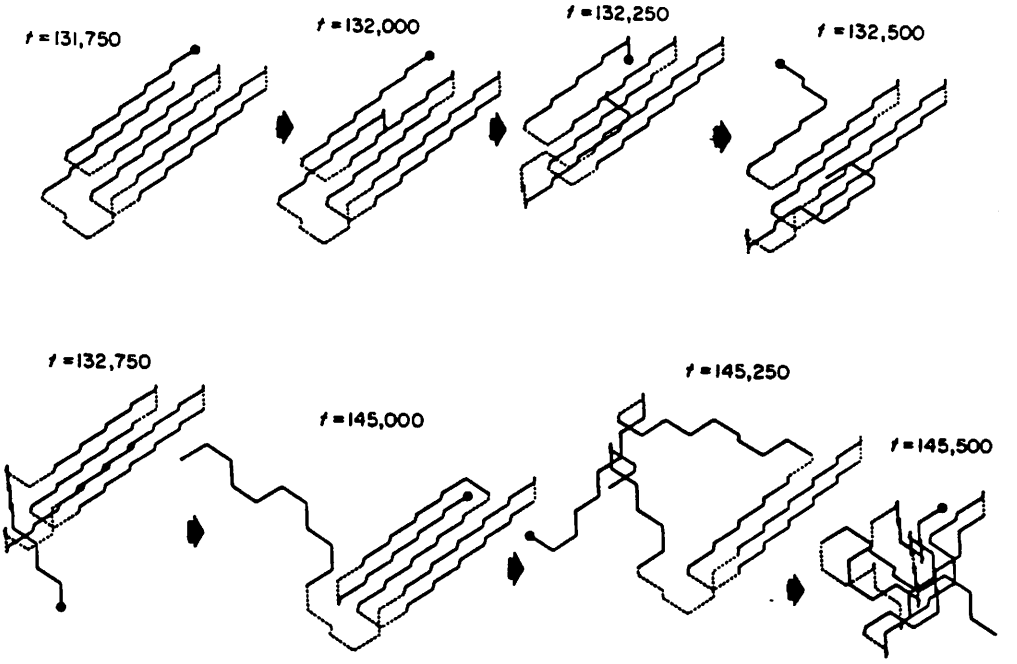


Figure 11. Representative unfolding trajectory from run 18 of model A. The circle denotes the N terminus.

into the conformation bearing the specified N_C native contacts but without any tertiary interactions. It is primarily, but not exclusively, entropic in origin. At $T^* = 1{\cdot}33$, observe that the term $A_{N_C} - E_{N_C} - A_D$ increases from 67·8 to 72·5 as N_C increases from 29 to 37; that is the entropy change from the transition state to the denatured state is 93·5% of the native state, whereas the energy changes from −59·625 to −70·875, or about 84% of the fully folded state. Thus, in agreement with the simulations presented by Taketomi *et al.* (1988), the entropy of the transition state and the native state are quite close.

The qualitative picture that emerges is as follows. Since the final transition state is very close to the fully folded structure, the difference between the transition state and the native free energy is primarily energetic; in this sense, the present simulations and the Cardboard Box model agree; whereas between I and the transition state, the difference in free energy is primarily entropic, (e.g. at $T^* = 1{\cdot}33$, the energy difference between $N_C = 26$ and 29 equals 4·885, whereas the entropic contribution to the free energy increase is 8·96).

Consider further the predicted behavior of refolding as a function of temperature. We focus on the height of the barrier between the four-member β-barrel intermediate and the native state. From $T^* = 1{\cdot}54$ to $T^* = 1{\cdot}18$, the barrier between I and the transition state ($A_{N_C} - A_D$ from $N_C = 26$ to 29) has changed from 4·44 to 3·78 at $T^* = 1{\cdot}18$; that is, a change of 35%, whereas the change in the free-energy barrier between $N_C = 29$ and $N_C = 37$ is 4·75 at $T^* = 1{\cdot}54$ and 8·43 at $T^* = 1{\cdot}18$. Thus, the rela-

Table 5

Compilation of parameters for the free energy calculation along the reaction co-ordinate

Total number of native contact pairs N_C				
Initial number of contact pairs	Final number of contact pairs	Representation of final structure assembled	Energy of assembled native section	Increase in number of rotational states in the native conformation
1	5		$16\varepsilon_c + 5\varepsilon_h - 6\varepsilon_g$	19
6	10		$16\varepsilon_c + 5\varepsilon_h - 6\varepsilon_g$	12
11	12		$32\varepsilon_c + 5\varepsilon_h - 6\varepsilon_g$	13
21	25		$16\varepsilon_c + 5\varepsilon_h - 6\varepsilon_g$	11
26	27		$8\varepsilon_h - 8\varepsilon_g$	7
28	37		$26\varepsilon_c + 20\varepsilon_h$	9

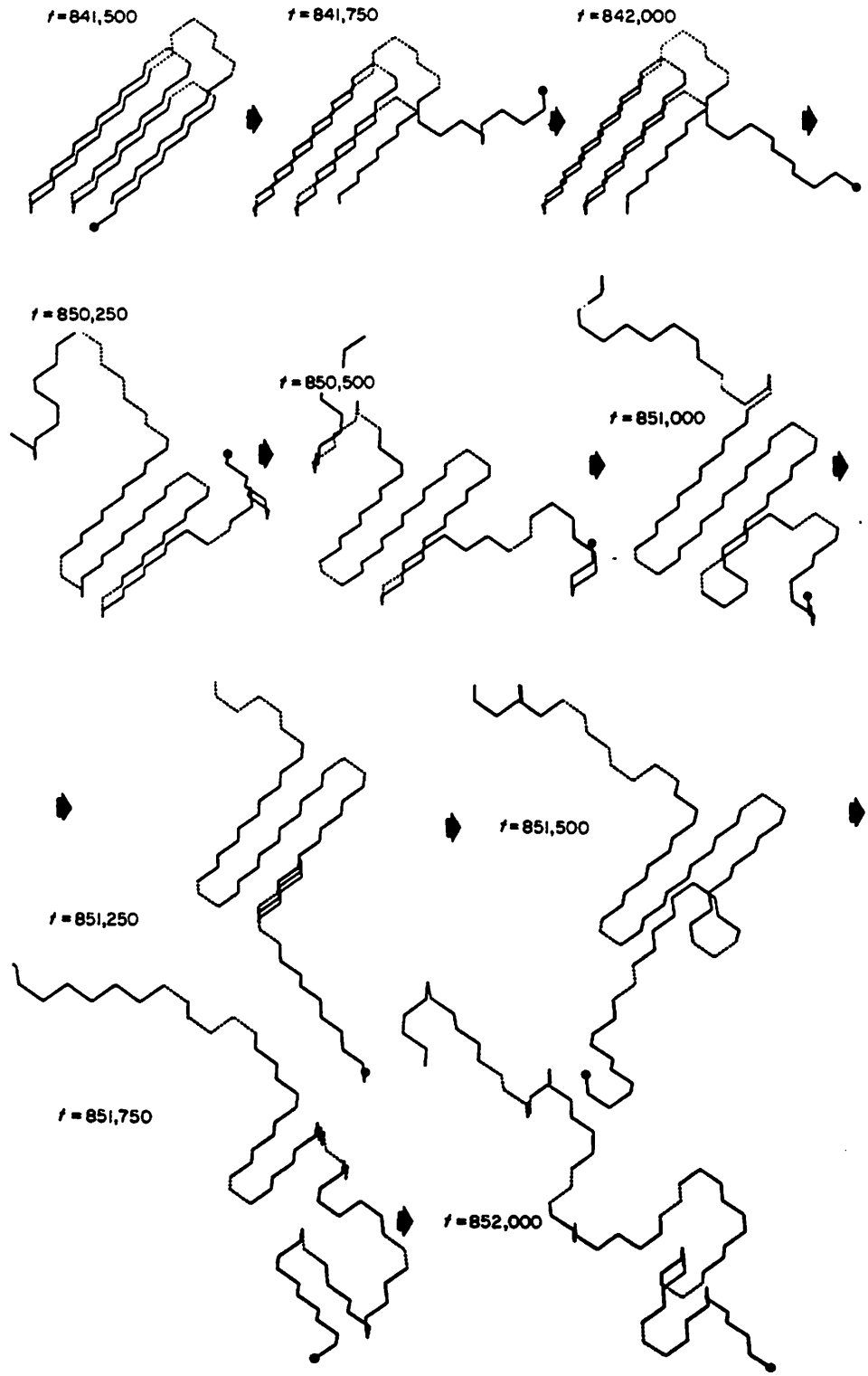


Figure 12. Representative unfolding trajectory from run 34 of model A. The circle denotes the N terminus.

tive rate of unfolding is expected to decrease far more rapidly (by about a factor of 40) than the rate of folding increases (by about a factor of 2) as the free energy of the native state has changed by 18·56 (in units of k_BT). Thus, as pointed out by Goldenberg & Creighton (1985), the free energy of the transition state relative to the native state changes less than the increase in stability of the native state. This is in qualitative agreement with experimental observations on different cytochromes *c* (Brems *et al.*, 1982) and different ribonuclease A conditions (Tsong & Baldwin, 1978), which indicate that the rate of folding is altered much less than the stability of the native state, thereby implying a greater decrease in the rate of unfolding. Thus, the punctuated assembly model developed here qualita-

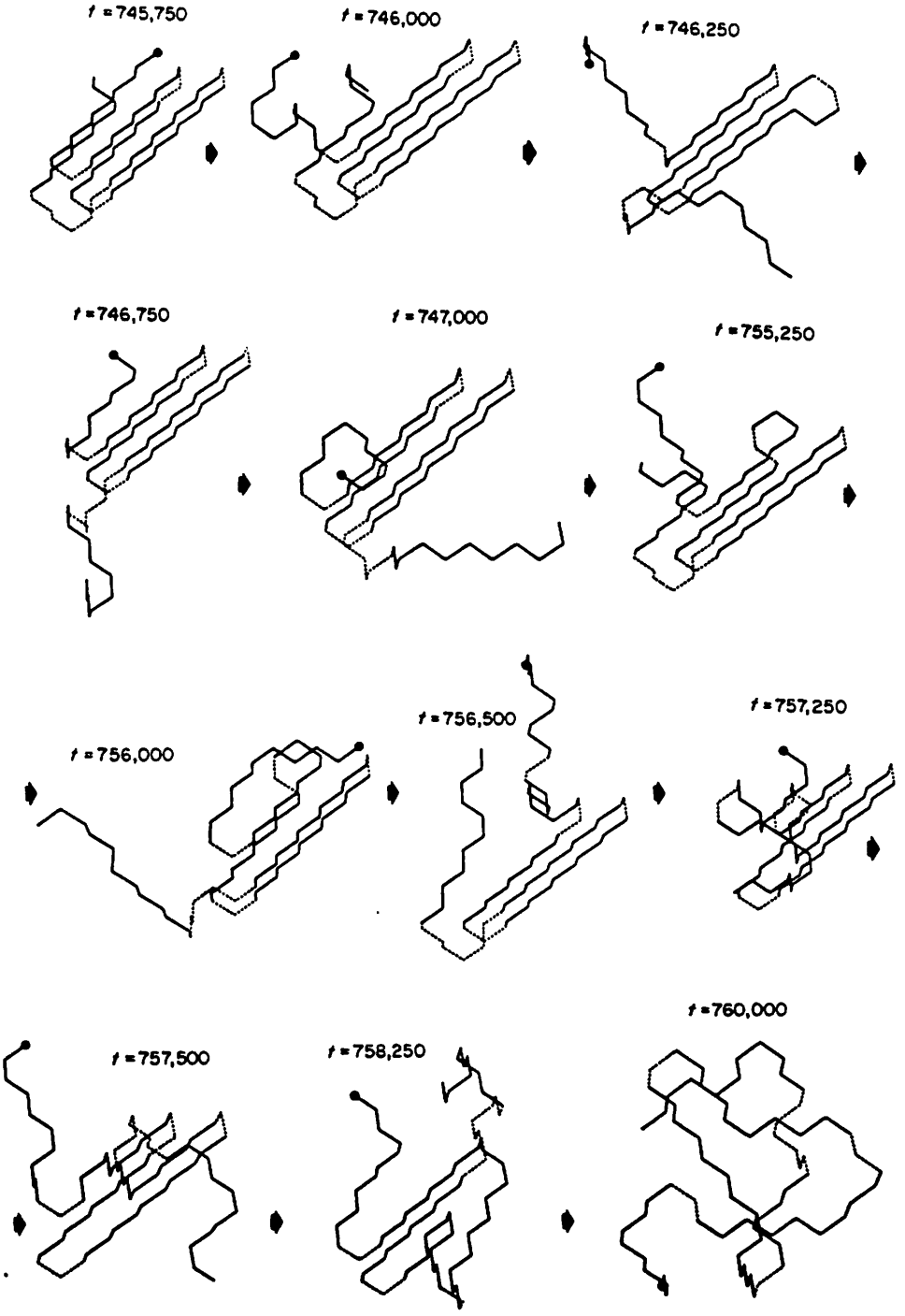


Figure 13. Representative unfolding trajectory obtained from run 9 of model B. The circle denotes the N terminus.

tively reproduces real experiments. (We point out that the qualitative features seen here hold for 4-helix bundles as well, and appear to be insensitive to the particular topology (Sikorski & Skolnick, 1990).)

Finally, we turn to the question of the relative rate of folding of model A compared to model B. Direct comparison at the same value of T^* is inappropriate, since model B has a higher transition temperature than model A; the two models should be compared under comparable conditions, for example at the transition midpoint. Moreover, the transition times compiled in Tables 3 and 4 are all approximate, and by no means exact. Thus, we employ the analytic theory embodied in equations (4) through (10). For model A, we find $T^*_{1/2} = 1{\cdot}237$ and for model B, $T^*_{1/2} = 1{\cdot}304$. In Figure 15, we plot $A_{N_C} - A_D$ for models A and B in the curves denoted by filled diamonds and squares respectively; observe $A_{37} - A_D = 0$ for both cases. For models A and B, the first barrier between D and the first intermediate state occurs at $N_C = 6$ and is of magnitude

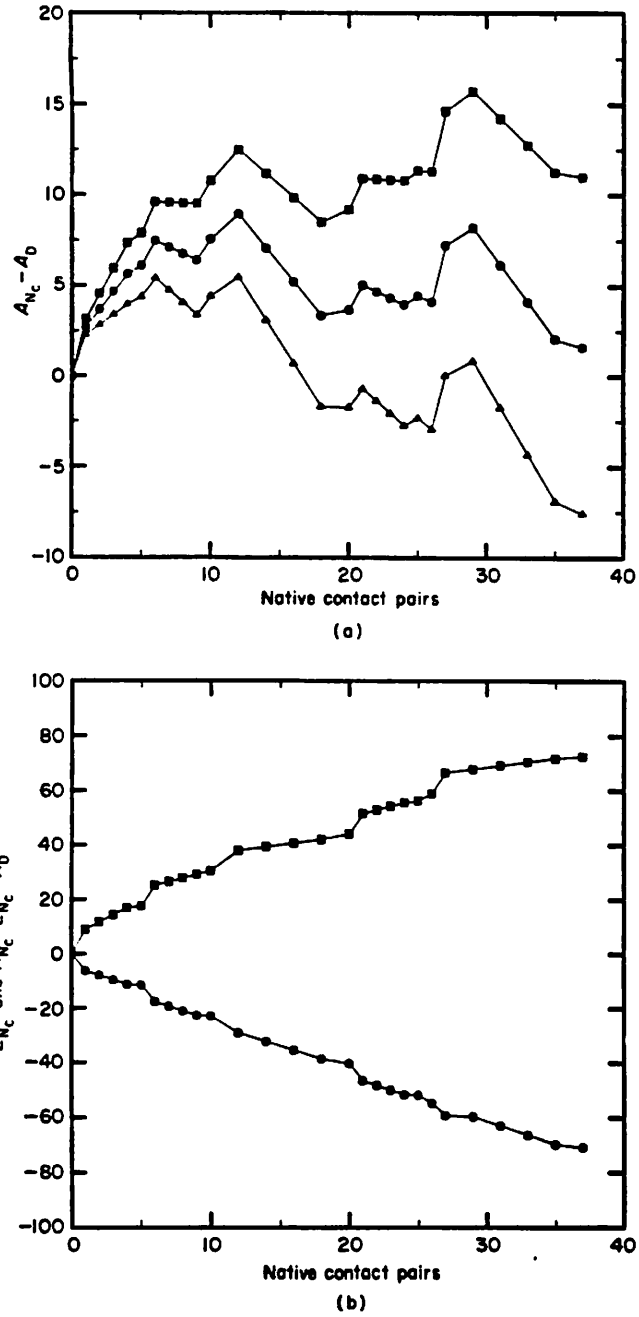


Figure 14. (a) Plot of $A_{N_C}-A_D$ *versus* N_C calculated *via* eqns (4) and (7) for model B at $T^* = 1{\cdot}54$, $1{\cdot}33$ and $1{\cdot}18$ in the curves denoted by filled squares, circles and triangles, respectively. (b) Plot of E_{N_C} (circles) and $A_{N_C}-A_D-E_{N_C}$ (the entropic contribution to the relative free energy difference: squares) *versus* N_C at $T^* = 1{\cdot}33$.

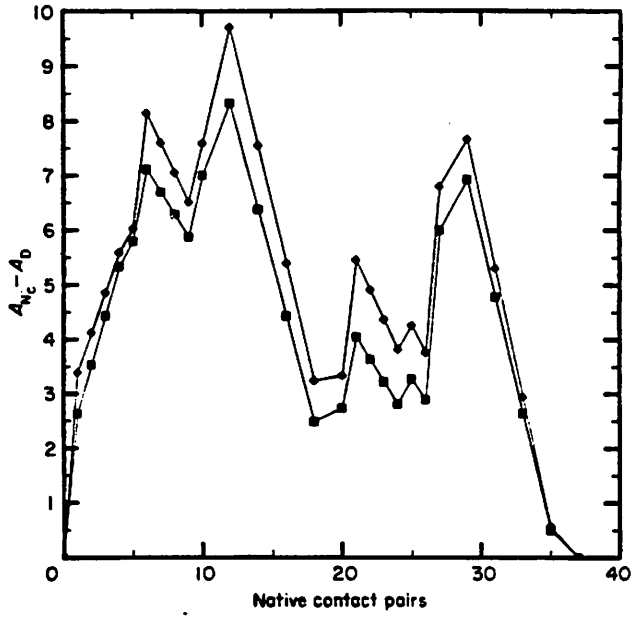


Figure 15. Plot of $A_{N_C}-A_D$ *versus* N_C at the transition midpoint for models A and B in the curves denoted by the filled diamonds and squares, respectively.

8·14 for model A and 7·12 for model B. Similarly, the height of the barrier at $N_C = 12$ from the local minimum $A_{N_D}-A_C$ at $N_C = 9$ equals 3·19 for model A and 2·44 for model B. Thus, because the formation of the four-member β-barrel intermediate involves the three triplets of energetically stabilized *gauche* states in model B as compared to model A, the folding intermediate is predicted to fold somewhat (but not more than an order of magnitude) faster in model B than in model A. This is reflected in the simulations by the more frequent appearance of the intermediate in model B (e.g. see Figs 7 and 8 *versus* 5). Otherwise stated, the intermediate in model B is more stable than model A; thus it should and was observed to form more frequently. We next consider the transition between I and N. Here, the energetics of the two models are identical, and effects due to the difference in transition temperatures of the two models should enter. Basically, because $T^*_{1/2}$ of model A is lower than that of model B, the configurational entropy term is less important in model A than it is in model B. The absolute magnitude of the transition state relative free energy (at $N_C = 29$) equals 7·66 in model A and 6·91 in model B. Thus, model B should unfold faster than model A, and the intermediate, being more stable, should be more populated; that is, transitions into the intermediate state from both D and N should be faster in model B than in model A. Qualitatively, both effects are borne out in the simulations (compare once again Figs 7 and 8 with 5). The transition-state free energy relative to the intermediate is best calculated between $N_C = 26$ and $N_C = 29$, and equals 3·90 for model A and 4·30 for model B. Because the barrier on the intermediate side is entropic in origin, the rate of folding should be slower for model B from the intermediate state to the native transition than it is for model A. Roughly speaking, Tables 3A and 4A indicate such a qualitative trend.

Overall, based on the comparison of models A and B, the stabilization of native-like turn conformations is seen to exert competing effects on the rate of

folding. For the formation of those sections that are directly stabilized, it accelerates the folding process. Moreover, a concomitant condition to the overall energetic stabilization of the native state is that the transition occurs at a higher temperature. Because the barrier between I and N is primarily entropic in these models, it slows down the transition from the intermediate to the native state. However, in all cases the effect is minor, less than an order of magnitude for the models considered here, even though the two native conformations differ in energy by about 15·6 (k_BT); the, at first, surprisingly small diminution in relative rates is due to the fact that the relevant free energy difference is between the native and partially folded conformations of the same molecule. Stabilizing turn-like conformations stabilizes both the native and denatured states, and it is their free energy difference that is important for the calculation of the free energy along the reaction co-ordinate.

4. Summary and Discussion

In the context of a highly simplified diamond lattice model of a six-member, Greek key β-barrel globular protein, the nature of the folding and unfolding pathways have been explored. However, in spite of the simplified nature of the model, it possesses the essential topological features of β-proteins, and consequently some conclusions applicable to real proteins can be drawn. Folding is seen to initiate close to a β-turn, followed by the zipping up of the adjacent β-strands on site. In the early stages of protein folding, the existing tertiary structure acts as scaffolding for subsequent formation of additional secondary structure that is assembled on site. Relatively rapid assembly of the four-member β-barrel intermediate occurs, after which there is a pause in assembly as the marginally stable intermediate hunts through configurational space to find the native state. In qualitative agreement with experiments (Creighton, 1985, 1988), the transition state is identified to be very close to the native state. In particular, it consists of β-strands 1 through 5 fully assembled, the long loop in the native conformation and the first native contact of the last strand. The transition state is an *almost fully assembled, but not distorted native state*. The folding intermediate is trapped in a broad free-energy valley that is separated from the transition state mainly by the reduction in configurational entropy relative to the denatured state. In this, the model disagrees with the Goldenberg–Creighton Cardboard Box model of protein folding. However, the free-energy barrier between the native state and the unfolded intermediate is primarily energetic in origin, in agreement with the Cardboard Box model. The predicted temperature-dependence of the unfolding and folding kinetics (Brems *et al.*, 1982; Tsong & Baldwin, 1978), is in qualitative agreement with experiment. Thus, these simulations support a framework model of protein folding. Finally, the unfolding pathway is basically the reverse of the folding pathway.

It is worthwhile to compare the punctuated on-site mechanism with some previous models of protein folding. In a number of realizations of the sequential assembly model (Kim & Baldwin, 1982), the secondary structure forms first, these then condense to form an intermediate globular structure, which then readjusts. In the present model, there is always fluctuating secondary structure; the tertiary structure is formed concurrently with secondary structure chacteristic of the native state. Because these models are on a low co-ordination number lattice, no subsequent local conformational readjustment is possible. While we cannot, on the basis of these simulations, rule out such local rearrangements, it is clearly not required to reproduce all the qualitative features of both the kinetics and thermodynamics of globular protein folding.

These simulations on β-barrel formation argue against a diffusion-collision-adhesion mechanism for folding initiation. Karplus & Weaver (1976, 1979) and Lee *et al.* (1987) have applied the model mainly to α-helical proteins; we discuss the validity of this model for the folding of α-helical proteins in the accompanying paper. Since β-sheets are less stable without tertiary interactions, no application has been made to β-proteins. The present models, having marginally stable β-strands without tertiary interactons, could assemble *via* what is effectively a diffusion-collision-adhesion mechanism. The reason they do not is that it is simply faster and more probable to initiate folding at or near a β-turn and then sequentially zip up rather than assemble the β-strands and wait for the strands to diffuse together before they dissolve.

In practice, it may be rather difficult to differentiate between the on-site construction model proposed here and a model involving preformation of secondary structure. Consider the case of a β-hairpin as an early folding intermediate. In both models, isolated β-strands of marginal stability are allowed. Thus, the initial (an isolated pair of strands) and final states (the β-hairpin) are identical. They differ only in what happens in between these two conformations. In one case, the isolated β-strands remain intact and diffuse together, in the other, perhaps one or both β-strands would dissolve, only to be reassembled at near or a turn with the β-strands formed by pulling pieces of the random coil tails from the denatured state. While it is certainly true that there are a number of artifacts present in this model (the degeneracy of mirror image conformers being the most striking example), these would tend to increase the number of folding and unfolding states by stabilizing non-physical intermediates. Thus, the fact that the folding pathway is rather well defined and, when coupled to a simple analytic theory, gives qualitative predictions in agreement with experiment, leads us to believe that the general mechanism of folding is physical. The differences in the folding sequence mainly occur in the early stages of protein folding,

where the free energies of the β-hairpin intermediates are essentially equal. While the general progression of events is better defined the closer one gets to the native state, there are a multiplicity of possible spatial trajectories consistent with the general course of folding events. Moreover, it shoud be pointed out that there is always a constant competition between tertiary structure assembly and dissolution. In more than one case, we have seen formation of β-strands 2, 3 and 4, which then dissolve because the long tail attached to strand 4 is too tangled to permit further formation of tertiary structure. The whole structure then dissolves and folding may or may not be reattempted soon thereafter.

In the accompanying paper (Sikorski & Skolnick, 1990), the nature of the folding and unfolding pathways of all the variants of left-handed, four-helix bundles is explored. We conclude that the punctuated on-site mechanism obtains even when the diffusion-collision-adhesion mechanism is expressly implemented as a competitive possibility. Thus, we believe the on-site construction mechanism is quite general, and we are applying the simulation technique to study naive models of *in vivo* folding as well as extending the model to a more realistic lattice description. Much work remains to be done.

This work was supported in part by NIH grant GM-37408 from the Division of General Medical Sciences, United States Public Health Service. Simulating conversations with Drs David Case, Alfred Holtzer, David Richardson, Jane Richardson, Andrzej Sikorski, and Peter Wright are gratefully acknowledged.

References

Anfinsen, C. B. (1972). *Biochem. J.* **128**, 737–749.

Baumgartner, A. (1984). *Annu. Rev. Phys. Chem.* **35**, 419–435.

Binder. K. (1984) Editor of *Application of the Monte Carlo Method in Statistical Physics*, chapt. 5, Springer, Berlin.

Binder, K. (1987) Editor of *Monte Carlo Method in Statistical Physics*, chapt. 1, Springer, Berlin.

Boots, H. & Deutch, J. M. (1977) *J. Chem. Phys.* **67**, 4608–4610.

Brandts, J. F., Halvorson, H. R. & Brennan, M. (1975). *Biochemistry*, **14**, 4953–4963.

Brems, D. N., Cass, R. & Stellwagen, E. (1982). *Biochemistry*, **21**, 1488–1493.

Bundi, A., Andreatta, R. H., Rittel, W. & Wüthrich, K. (1976). *FEBS Letters*, **64**, 126–129.

Bundi, A., Andreatta, R. H. & Wüthrich, K. (1978). *Eur. J. Biochem.* **91**, 201–208.

Chan, H. S. & Dill, R. A. (1989). *Macromolecules*, in the press.

Chandrasekhar, S. (1943). *Rev. Mod. Phys.* **15**, 1–89.

Creighton, T. E. (1981). In *Structural Aspect of Recognition and Assembly in Biological Macromolecules* (Balaban, M., Sussman, J. L., Traub, W. & Yonath, A., eds), pp. 57–73, Balaban ISS, Rehovot.

Creighton, T. E. (1985). *J. Phys. Chem.* **89**, 2452–2459.

Creighton, T. E. (1988). *Proc. Nat. Acad. Sci., U.S.A.* **85**, 5082–5086.

Dubois-Violette, E., Geny, F., Monnerie, L. & Parodi, O. (1969). *J. Chim. Phys.* **66**, 1865–1871.

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, (1969). *Statistical Mechanics of Chain Molecules*, Wiley, New York.

Garel, J. R. & Baldwin. R. L. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 3347–3351.

Ghelis, C. & Yon, J. (1982). *Protein Folding*, Academic Press, Orlando, FL.

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.

Harrison, S. C. & Durbin, R. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 4028–4030.

Hill, T. L. (1956). *Statistical Mechanics*, chapt. 6, McGraw-Hill, New York.

Iwata, K. & Kurata, M. (1969). *J. Chem. Phys.* **50**, 4008–4013.

Karplus, M. & Weaver, D. L. (1976). *Nature (London)*, **160**, 404–406.

Karplus, M. & Weaver, D. L. (1979). *Biopolymers*, **18**, 1421–1427.

Kawajima, K., Hiraoka, Y., Ikeguchi, M. & Sugai. S. (1985). *Biochemistry*, **24**, 874–881.

Kim, P. S. & Baldwin, R. L. (1982) *Annu. Rev. Biochem.* **51**, 459–489.

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*a*). *Biopolymers*, **26**, 937–962.

Kolinski, A., Skolnick, J. & Yaris, R. (1987*b*). *J. Chem. Phys.* **86**, 1567–1585.

Kremer, K., Baumgartner, A. & Binder, K. J. (1981). *Phys. A.* **15**, 2879–2892.

Krigbaum, W. R. & Lin, S. F. (1982). *Macromolecules*, **15**, 1135–1145.

Lee, S., Karplus, M., Bashford, D. & Weaver, D. (1987). *Biopolymers*, **26**, 481–506.

Lesk, A. M. & Rose, G. D. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4304–4308.

Levinthal, C. (1968). *J. Chim. Phys.* **65**, 44–45.

Meirovitch, M., Vasques, M. & Scheraga, H. A. (1988). *Biopolymers*, **27**, 1189–1204.

Miyazawa, & Jernigan, (1982). *Biopolymers*, **21**, 1333–1363.

Oas, T. G. & Kim P. S. (1988). *Nature (London)*, **336**, 42–48.

Privalov, P. L. (1979). *Advan. Protein Chem.* **33**, 167–241.

Ptitsyn, O. B. & Funkelstein, (1980). *Quart. Rev. Biophys.* **13**, 339–386.

Ptitsyn, O. B. & Rashin, A. A. (1975). *Biophys. Chem.* **3**, 1–28.

Richardson, J. S. (1981). *Advan. Protein Chem.* **34**, 167–339.

Scheraga, H. A. (1973). In *Current Topics in Biochemistry* (Anfinsen, C. B. & Schechter, A. N., eds), pp. 1–42, Academic Press, New York.

Scheraga, H. A. (1980). In *Protein Folding* (Jaenicke, R., ed.), pp. 261–288, Elsevier/North Holland, Amsterdam.

Shoemaker, K. R., Kim, P. S., Brems, D. N., Marqusee, S., York, E. J., Chaikin, I. M., Stewart, J. M. & Baldwin, R. L. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 2349–2353.
Shoemaker, K. R., Kim, P. S., York, E. J., Stewart, J. M. & Baldwin, R. L. (1987). *Nature (London)*, **326**, 563–567.
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.
Sikorski, A. & Skolnick, J. (1990). *J. Mol. Biol.* **212**, 819–836.
Skolnick, J. (1983). *Macromolecules*, **16**, 1763–1770.
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.
Teramoto, E., Kurata, M. & Yamakawa, H. J. (1958). *J. Chem. Phys.* **28**, 785–791.
Tsong, T. Y. & Baldwin, R. L. (1978). *Biopolymers*, **17**, 1669–1678.
Udgaonkar, J. B. & Baldwin, R. L. (1988). *Nature (London)*, **335**, 694–699.
Ueda, Y. Taketomi, H. & Gō, N. (1978), *Biopolymers*, **17**, 1531–1548.
Valeur, B. Jarry, J. P., Geny, F. & Monnerie, L. (1975). *J. Poly. Sci. Poly. Phys. Ed.* **13**, 667–674.
Weaver, D. L. (1984). *Biopolymers*, **23**, 675–694.
Wetlaufer, D. B. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 697–701.
Wright, P. E., Dyson, H. J. & Lerner, R. H. (1988). *Biochemistry*, **27**, 7167–7175.
Zielenkiewicz, P. & Rabczenko, A. (1988). *Biophys. Chem.* **29**, 219–224.

Edited by T. J. Richmond

Note added in proof. Recently, we undertook a series of dynamic Monte Carlo simulations on a 24 nearest-neighbor lattice representation of proteins in which both finite backbone chain thickness and side-chains are included. The folding and unfolding pathways of a full 99 residue analogue of plastocyanin have been simulated. The major folding pathway is qualitatively identical with those reported here with the exception that strands 3,4 serve as the dominant initiation site. Thus, these pathways appear to be universal; that is, they are independent of the particular lattice realization and model details as well as the choice of local elemental moves.