Diffusion-Collision Model Study of Misfolding in a Four-Helix Bundle Protein

Chris Beck, Xavier Siemens, and David L. Weaver
Molecular Modeling Laboratory, Department of Physics, Tufts University, Medford, Massachussetts 02155 USA

ABSTRACT Proteins with complex folding kinetics will be susceptible to misfolding at some stage in the folding process. We simulate this problem by using the diffusion-collision model to study non-native kinetic intermediate misfolding in a four-helix bundle protein. We find a limit on the size of the pairwise hydrophobic area loss in non-native intermediates, such that burying above this limit creates long-lasting non-native kinetic intermediates that would disrupt folding and prevent formation of the native state. Our study of misfolding suggests a method for limiting the production of misfolded kinetic intermediates for helical proteins and could, perhaps, lead to more efficient production of proteins in bulk.

INTRODUCTION

Proteins fold spontaneously in living systems into many different functional forms, the number of types approaching 10^5 in humans. These proteins control most events in our biological lives. The end result of folding, the final folded structure, is called the native structure. It determines the behavior of the proteins and consequently their particular role in a living system. The human genome project promises to provide the amino acid sequences of all of these proteins in the next few years, and it will be very important to have their native structures available at atomic resolution, as well. Because of the very large number of native structures to be determined, it will be essential to have significant theoretical help in predicting structures. An important theoretical approach in folding studies is to try to simulate the kinetics of the folding process from amino acid sequence to the final native structure fold. For example, a 36-residue three-helix bundle protein, the villin headpiece subdomain, was recently the object of an attempt to determine the folding pathway by a molecular dynamics simulation in explicit solvent (Duan and Kollman, 1998). A single 1-μs trajectory was calculated and it ended in a relatively stable structure (lifetime of ~150 ns), which was suggested to be a folding intermediate. Although attempts to overcome the computational bottleneck for direct simulations by use of a very large network of PCs are in progress, it is likely that alternatives approaches, based on simplified models, will be of importance for some time to come. Moreover, even when brute-force folding has been achieved for some proteins, the interpretation of the results is likely to be made with such models. The diffusion-collision model (Karplus and Weaver, 1976, 1994) is a model that has been applied successfully to a number of helical proteins, including apomyoglobin (Pappu and Weaver, 1998), the study of a series of mutants of the monomeric λ-repressor (Burton et al., 1998), and several three-helix bundle proteins (Islam et al., submitted for publication). Alternatively, efforts are being made to predict the native structure from the sequence using information theory methods that do not necessarily involve the physics of the folding process. Given much information about homologs, the latter method can be quite successful. However, there may be sequences in the human genome database for which no homologs exist and for which a more kinetic physical approach will be useful.

In folding kinetics, there is randomness associated with the behavior of individually folding molecules due to two factors. First, if there are k folding elements, then there are $(k(k-1))/2$ possible pairings among them. Not all of the possible pairings are between elements that are paired in the native structure (native pairings). Second, propensities in amino acid sequences among α-helix, β-strand, and other local structures can be influenced by long-range interactions. For example, helical regions in the native structure could transiently be β-strand in early folding kinetic events (see, for example, Fezoui et al., 2000).

It is important for functional proteins to avoid the consequences of random folding because it can lead to misfolding of several types. Any misfolding, both for biomedical (protein misfolding diseases) and bioeconomic (formation of misfolded aggregates in large-scale production) reasons is to be avoided. Understanding the molecular mechanisms of unimolecular and bimolecular misfolding may lead to advances in biomedicine and in protein production improvements. In fact, in protein misfolding diseases, both types of random behavior are apparent, with an $\alpha \rightarrow \beta$ transition in a folding element, followed by aggregation (see, for example, Fezoui et al., 2000 for a possible misfolding model with both elements).

In this paper we will concentrate on the first type of random misfolding, namely that involving non-native pairing of folding elements (Ikai and Tanford, 1973; Kho-

Received for publication 15 March 2001 and in final form 6 August 2001.
Address reprint requests to Dr. David Weaver, Molecular Modeling Laboratory, Tufts University, Medford, MA 02155. Tel.: 617-627-3515; Fax: 617-627-3878; E-mail: dweaver@tufts.edu.

0006-3495/01/12/3105/11 $2.00

rasanizadeh et al., 1996; Baldwin, 1996) and, in particular, on the (mis)folding of four-helix bundle proteins.

The kinetics of folding of four-helix bundle proteins has been widely studied in recent years with evidence of different kinetic behavior in homologous proteins. Kragelund et al. (1995, 1996, 1999) have investigated the folding of the family of four-helix bundle proteins that bind medium and long-chain acyl-coenzyme A esters with very high affinity. The known three-dimensional structures are very similar, although there can be substantial differences in individual amino acid sequences. Folding and unfolding rates in water can also differ by more than an order of magnitude. Ferguson et al. (1999) have studied the kinetics and thermodynamics of folding of the homologous four-helix bundle proteins Im7 and Im9, with Im9 being an apparent two-state folder (Jackson, 1998) and Im7 folding via an on-pathway intermediate (Capaldi et al., 2001). These proteins, which inhibit the cytotoxic activity of the E colicin DNase proteins, preventing the death of the colicin-producing bacteria (Wallis et al., 1992), have 60% sequence homology and have the same three-dimensional structure. Nevertheless, they appear to fold by different kinetic mechanisms. Diffusion-collision model principles (Karplus and Weaver, 1976, 1979, 1994) could help elucidate the similarities and differences in the kinetic behavior of these families. We will use the diffusion-collision model to study the native and non-native folding kinetics of four-helix bundle proteins.

The diffusion-collision model (Karplus and Weaver, 1976, 1979, 1994) views the kinetic folding process of an unfolded α-helical protein as the diffusion, collision, and coalescence of marginally stable helices, called microdomains. In previous applications of the model, only native collision-coalescence events have been considered (Bashford, et al., 1984, 1988; Yapa and Weaver, 1992, 1996; Pappu and Weaver, 1998; Burton et al., 1998; Rojnuckarin et al., 1998; Vasilkoski and Weaver, 2000; Islam et al., submitted for publication). In this work we consider all of the possible helix-helix pairings in a four-helix bundle protein and assess the significance of non-native collisions and transient coalescence events on folding to the native structure. The diffusion-collision model does not impose a particular kinetic mechanism on protein folding. According to the model, a protein could fold faster or slower, by one or several significant paths and with or without apparent intermediates; the folding properties depending mainly on the microdomains, their amino acid sequences, and their chain distances from one another, which can be changed by mutation or genetic engineering.

In a four-helix bundle protein there are four α-helices packed together in one of the standard four-helix bundle motifs (Weber and Salemme, 1980; Sheridan, et al., 1982; Paliakasis and Kokkinidis, 1992) or in less regular arrangements. The helices are labeled A–D starting at the N-terminus and there are six possible pairings among them, namely AB, AC, AD, BC, BD, and CD. The number of actual helix-helix pairings in a native structure may be fewer than the maximum number possible. It depends on the three-dimensional packing geometry of the four helices, their lengths, and the lengths of the connecting loops, with some pairings forbidden by steric hindrance. There could, in principle, be three, four, five, or six pairings involving all four helices. The complexity of diffusion-collision model calculations depends on the number of helix-helix pairings n, with the number of folding states being 2^n, the number of transitions among the states being $n2^{n-1}$ and the number of independent pathways from the no-pairs state to the all-pairs state being $n!$. The possibilities are shown in Table 1 for between three and six pairings. As the number of native pairings is reduced, the possibility of encountering transient non-native pairings increases. Clearly, introducing non-native pairings as possible collision, transient coalescence candidates will increase the complexity of folding by increasing the number of kinetic intermediate states. Intermediate states can either speed up or slow down folding, depending on their energy and their ease of progressing toward the final state (Khorasanizadeh et al., 1996; Jackson, 1998; Wagner and Kiefhaber, 1999). Kinetic intermediates in deep wells will reduce the probability fraction in the final folded state and increase the overall folding time.

TABLE 1 Four-helix bundle protein possible pairings

No. Native Pairings n	No. Native States 2^n	No. of Possible Transitions $n!$	No. of Independent Pathways $n2^{n-1}$
3	8	12	6
4	16	32	24
5	32	80	120
6	64	192	720

In the diffusion-collision model, if a four-helix bundle protein is fully paired with six helix-helix pairings, there is no chance for non-native pairings among the helices of the kind being considered here. With five or fewer pairings in the native structure, non-native pairings may occur in the kinetic processes of the model. Looking at Table 1, we see that with five native pairings, there will be 32 possible non-native kinetic states; with four native pairings, there will be 48 possible non-native kinetic states; and with three native pairings, there will be 56 possible non-native kinetic states. The possibilities of pairing among helices in the native structure may be illustrated by schematically representing the helices as right-circular cylinders of the same length. If the four helices have their cylinder axes approximately parallel or antiparallel, as shown from a top view in Fig. 1, *a* and *b*, then there are four possibilities for packing. Fig. 1 *c* shows square packing, with each helix being in contact with its two cyclic neighbors. With this packing arrangement there are four helix-helix pairings (AB, BC, CD,

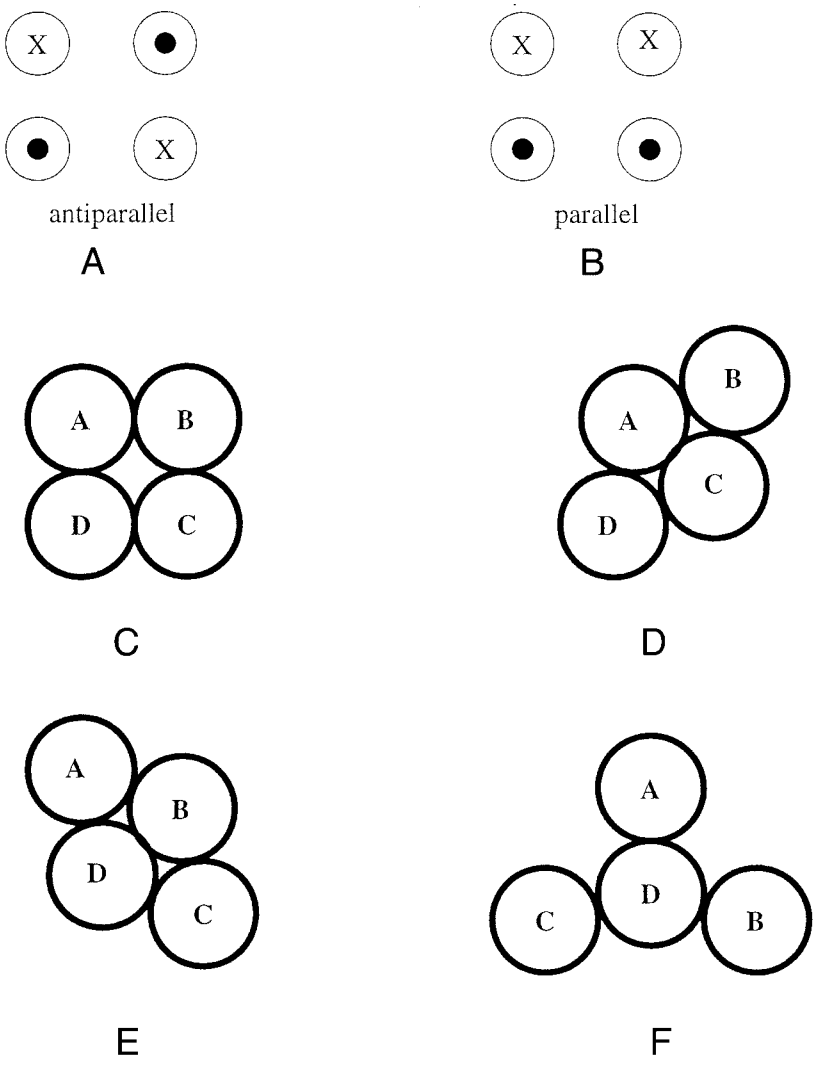

FIGURE 1 Schematic diagrams (top view) showing possible packing arrangements in idealized four-helix bundles. In *a* and *b*, an X shows the N-terminal of a helix going into the paper; ● shows the C-terminal end of a helix coming out of the paper. The helices are ordered in a clockwise manner. Panel *c* shows the square arrangement of the four helices A–D, which permits four helix-helix pairings (AB, BC, CD, AD). Panels *d* and *e* show the two diagonal arrangements of the four helices, which permits five helix-helix pairings AB, BC, CD, AD and either AC (*d*) or BD (*e*). Panel *f* shows an arrangement of the four helices that has three helix-helix pairings.

and AD) and 16 native kinetic states. Fig. 1, *d* and *e* show the shifting of one side or the other of the Fig. 1 *c* square packing to make diagonal packing. Here, a diagonal pair of helices also make hydrophobic contact, either AC (Fig. 1 *d*) or BD (Fig. 1 *e*). In both cases there are five helix-helix pairings, the four native pairings shown in Fig. 1 *c* and one diagonal (AC or BD) pairing and 32 kinetic states. A three-pair four-helix bundle protein with eight kinetic states could be formed by having three of the helices form an approximate equilateral triangle about the remaining helix, for example, see Fig. 1 *f* with the A, B, and C helices surrounding the D helix. In real proteins, there are at least three possible deviations from ideal behavior: 1) the helical axes are not aligned; 2) the helices are different lengths; and 3) the loops between helices may be short or long and may be different lengths. For example, cytochrome b562 (PDB code 256b) has an approximately antiparallel arrangement of the four helices (see Fig. 1 *a* and Fig. 2). The helix lengths are A (17 res.), B (18 res.), C (25 res.), and D (22 res.) and the three loops have lengths 3, 15, and 3 residues, respectively. The helix axes are at small angles with respect to one another. There are three major pairings AB (828 $Å^2$ of solvent-accessible area loss upon packing in the native structure), BC (915 $Å^2$), and CD (990 $Å^2$). The other cyclic pairing AD has a smaller but still substantial area loss of 461 $Å^2$. The diagonal pairing BD (325 $Å^2$) has a smaller area loss and AC (203 $Å^2$) has an even smaller area loss. Cytochrome b562 appears to be a four-helix bundle protein with six helix-helix pairings consisting of

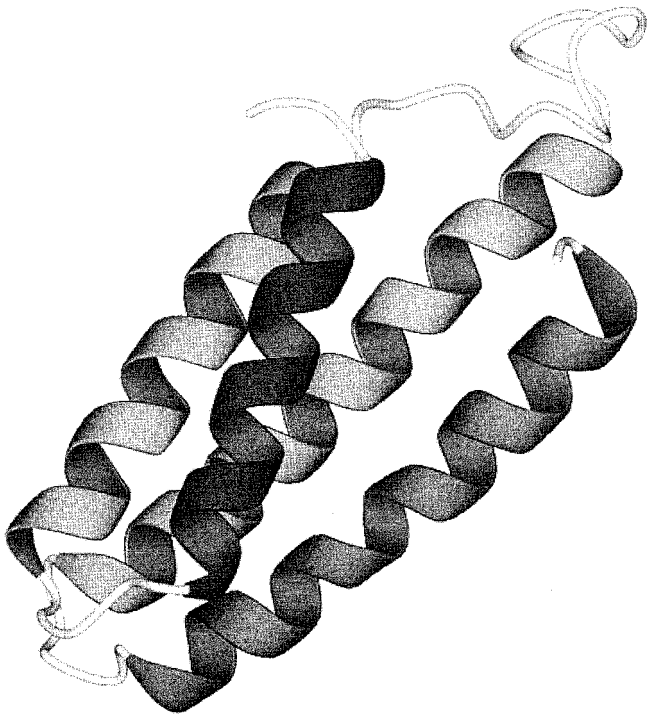

FIGURE 2 Cartoon representation of the cytochrome b562 crystal structure (PDB code 256b). The helix-helix pairings are described in the text. The program PREPI was used to make the drawing.

four very strong pairings (AB, BC, CD, and AD), one (BD) moderately strong pairing, and one (AC) rather weak pairing.

We use as our four-helix bundle protein an idealized model protein based on the work of Regan and DeGrado (1988). We previously built a model of this engineered protein using the structures of hemerythrin and myohemerythrin (Yapa and Weaver, 1992, 1996) as guides. In the present paper we will use a symmetrized model structure based on the Regan-DeGrado (1988) sequence. The symmetrized model four-helix bundle protein has all helices and loops identical in length and amino acid composition to the Regan-Degrado (1988) sequence. We assume that all helix axes are aligned and all cyclic solvent-accessible area losses upon helix-helix pairing (AB, BC, CD, AD) are equal and equal to 600 $Å^2$ per pairing. We will use the idealized model structure to examine diffusion-collision model folding with non-native intermediates. We show that diffusion-collision model calculations easily include non-native pairings and consequently non-native intermediates in folding simulations. We find that inclusion of non-native intermediates suggests an important limit on the stability of non-native pairings and their ability to be included in the folding kinetics of a protein without disrupting the final native fold.

The diffusion-collision model proposes that the folding process for a four-helix bundle protein be divided into a sequence of random helix-helix collisions, some of which result in coalescence of the helices. If all six helix-helix pairing are made, there are 64 possible diffusion-collision native kinetic states for this protein, shown schematically in Fig. 3. The 64 states in Fig. 3 are shown as boxes, with allowed transitions shown as arrows leading to and from connecting states. In the Fig. 3 kinetic states diagram, the kinetic states for three native helix-helix pairings are states 1-8, the kinetic states for four native helix-helix pairings are states 1-16, and the kinetic states for five native helix-helix pairings are states 1-32. The decimal label of a kinetic state is the binary number in base 10 plus one. For example, state 1 has no pairings, the decimal equivalent of binary (000000) plus one and state 64 has all six possible pairings, the decimal equivalent of binary (111111) plus one.

As outlined in Methods, diffusion-collision model calculations involve computing the rate matrix for the set of transitions (*arrows* in Fig. 3) between kinetic states. Each transition has a forward τ_f^{-1} and backward τ_b^{-1} rate. In prior diffusion-collision calculations only native helix-helix packing arrangements have been used for the kinetic intermediates. Although it is known that non-native collisions among helix pairs must occur randomly, it has been assumed in prior work that coalescence of non-native pairs is transient enough that it does not significantly affect the folding kinetics of the natively paired states. Our intent in these simulations is to assess the sensitivity of folding kinetics to this assumption. We do this by varying the solvent-accessible surface area buried upon coalescence of a non-native pairing of helices (AC, BD). Solvent-accessible area loss affects both the folding time through the parameter β, which includes an orientational contribution depending on the relative solvent-accessible area loss upon coalescence, and the unfolding time through the solvent-accessible area loss term ΔA associated with attractive hydrophobic interaction of a helix pair (see Methods for details). Hydrophobic area loss is known to be a principal factor in protein stability and is particularly important in all-helical proteins. Relatively small area loss upon pairing has been used in prior diffusion-collision simulations to differentiate among pairings that are more or less significant in the folding kinetics. In those studies we generally did not include the area loss contribution to β mentioned above. Rather, we assumed that once an association had occurred between helices, rearrangement to the native orientation was rapid compared to other time scales involved in coalescence or that most of the solvent-accessible surface area of a microdomain was available initially, making the contribution to β close to unity. However, in this study we explicitly include orientational effects, using the fractional area loss upon folding for each microdomain in a pairing as a multiplicative factor in β. This makes the folding rate of a pairing more sensitive to area loss.

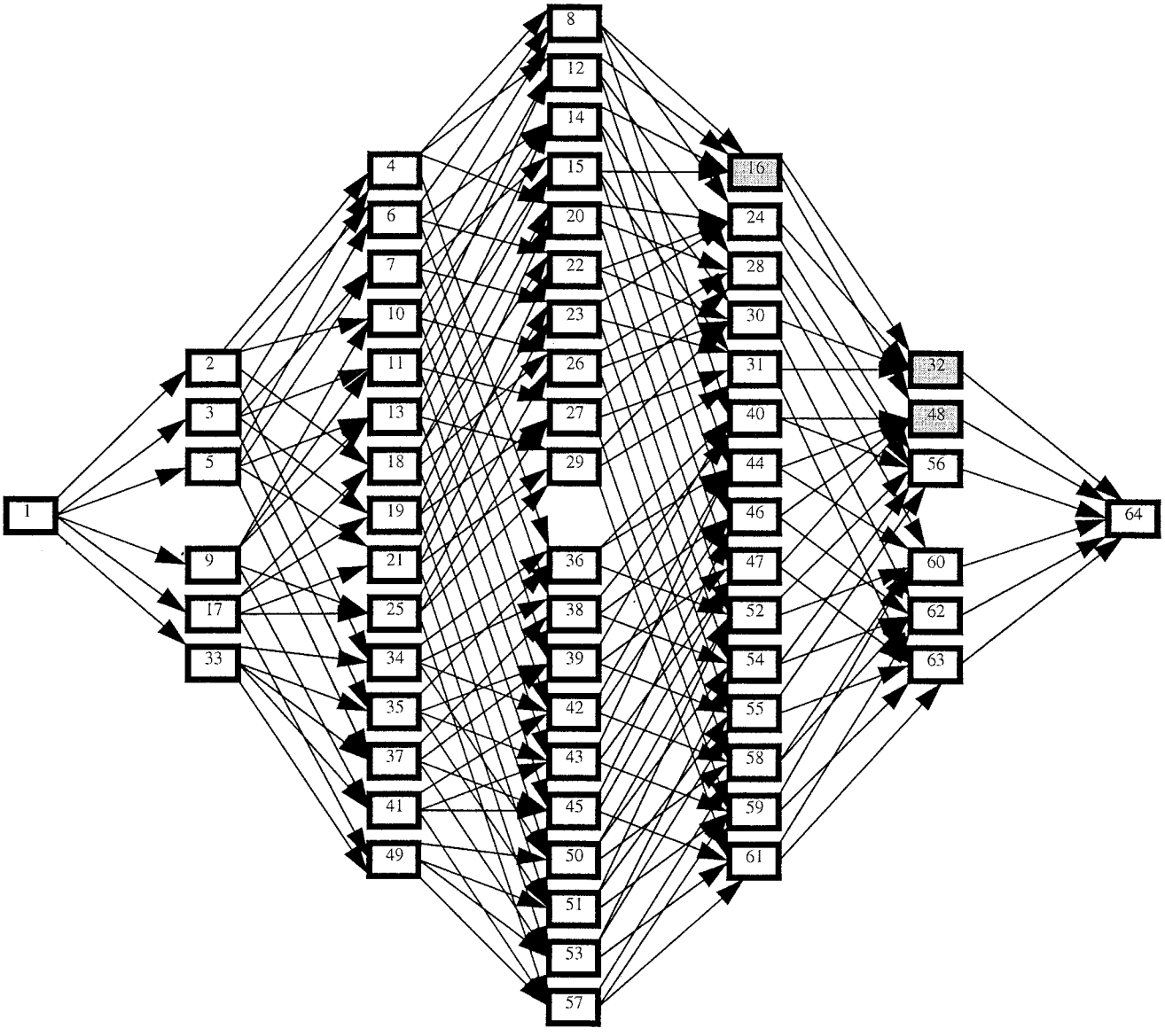

FIGURE 3 Diffusion-collision state diagram showing the kinetic states and pathways for the diffusion-collision folding of a four-helix bundle protein with six helix-helix pairings and 64 states. On the state diagram, the kinetic states for only three helix-helix pairings are states 1-8; the kinetic states for four helix-helix pairings are states 1-16; and the kinetic states for five helix-helix pairings are states 1-32.

In Results we present the results of several diffusion-collision model simulations of folding in the idealized four-helix bundle protein (Regan and DeGrado, 1988; Yapa and Weaver, 1992, 1996) in which non-native pairing of helices causes misfolding to a greater or lesser extent. We find that non-native kinetic intermediates that have sufficient buried hydrophobic area will not unfold or dissociate easily enough, and therefore will cause the protein to misfold. The exact area loss to cause misfolding depends on several factors, but a reasonable estimate for the model under study may be found by comparing the folding rates for non-native kinetic states and the native state. For early intermediates having fewer pairings than the native protein, the non-native intermediate must simply have a higher rate of unfolding than folding. This ensures that probability moves out of the misfolded state toward native states. The unfolding time of the non-native intermediate will affect the overall folding time, slowing it considerably if it is fairly stable. However, for intermediates with more pairings than the native state, the rate of folding from the native state to the non-native intermediate must be several times slower than the unfolding rate. This ensures that at equilibrium the native state is best described by the smaller number of pairings. The results are followed by a summary and discussion of implications of non-native kinetic interme-

diates for folding. The next section summarizes some details of diffusion-collision model calculations.

METHODS

The folding time is given by

$$\tau_f = \frac{l^2}{D} + \frac{L\Delta V(1-\beta)}{\beta DA} \quad (1)$$

The derivation of Eq. 1 has been given previously (Bashford et al., 1988; Karplus and Weaver, 1994). The volume available for diffusion of each microdomain pair ΔV, their relative target surface area for collisions A, their relative diffusion coefficient D, and their relative geometry parameter l^2 are calculated for diffusion in a spherical space. The parameter β is the product of four terms (each less than or equal to unity), two folding β's and two orientational β's. The Regan-DeGrado helical sequence has a folding β (from AGADIR) of almost unity (sequence tends to be helical), so the major contributors to β are the orientational β's found by dividing the solvent-accessible area loss for a microdomain in a particular microdomain pair by the total solvent-accessible area of the microdomain or cluster.

The unfolding time is given by

$$\tau_b = R_w\left(\frac{8\pi\mu}{k_B T}\right)^{1/2} e^{(f\Delta A)/(k_B T)} \quad (2)$$

This is an extension of the previously used unfolding time (Bashford et al., 1988; Karplus and Weaver, 1994) in which the attempt rate has been made microdomain pair-specific (Beck and Siemens, unpublished results). The parameter R_w is the width of the interaction space of two coalesced microdomains, taken to be the size of a water molecule for the hydrophobic interaction. The parameter μ is the reduced mass of the microdomain pair involved in unfolding. The parameter ΔA is the total solvent-accessible area loss upon pairing by the particular microdomain pair and f is the stabilization energy per unit area loss (Chothia, 1974). Equation 2 assumes that the unfolding path is linear.

The rates are used in the rate matrix formed from the coupled unimolecular pairing and unpairing processes (see, for example, Appendix A in Karplus and Weaver, 1994). Numerical solution of the rate problem provides the probability as a function of time of each of the possible pairing combinations.

RESULTS

The folding kinetics of the designed four-helix bundle protein (Regan and DeGrado, 1988), previously studied using native helix-helix pairings (Yapa and Weaver, 1992, 1996), was the starting point of our simulations. The Regan-DeGrado sequence consists of four α-helices, each with amino acid sequence GELEELLKKLKELLKG, connected by three loops, each with sequence PRR. An AGADIR calculation of the helical propensity (Munoz and Serrano, 1994a–c, 1997; Lacroix et al., 1998) of the helical sequence using the EMBL on-line calculation tool found 75% helicity at 293 K, pH 7, and ionic strength 0.1, with acetylated N-terminus and amidated C-terminus. Helical propensity enters the folding time through the parameter β. The helical propensities used in the simulations are larger than those found in naturally occurring four-helix bundle proteins using AGADIR estimates (results not shown). A larger β leads to a shorter folding time, so the times found in the simulations are one or two orders of magnitude smaller than found in naturally occurring four-helix bundle proteins (Kragelund et al., 1995, 1996, 1999; Jackson, 1998; Ferguson et al., 1999; Capaldi et al., 2001). With short loops and identical helices, the possible packing arrangements of the four helices are like Fig. 1, *c*, *d*, or *e*, that is, either square (Fig. 1 *c*) or diagonal (Fig. 1, *d* and *e*) packing. In addition, if helix-helix pairing AC forms then pairing BD is sterically prohibited, and vice-versa. In the diffusion-collision model, to account for the non-native helix-helix pairing steric effects shown in Fig. 1, *d* and *e*, we set the folding rates to zero in the rate matrix to form the other pairing (e.g., BD) when one of a sterically forbidden pairing has been made (e.g., AC). Thus, the BD pairing cannot form if the AC pairing has formed and vice versa, as seen Fig. 1, *d* and *e*. Forming one pairing blocks the possibility of the other pairing. In other four-helix bundle proteins, where the interhelical loops are longer, the helices are of different lengths and the helical orientations less regular, the sixth pairing may be made.

To examine the effect of non-native helix-helix pairings on the folding of a four-helix bundle protein, we performed a set of simulations with different values for the helix-helix solvent-accessible area losses of the AC and BD helix-helix pairings, keeping the cyclic pairing area losses fixed, as summarized in Table 2. We chose area loss values for the cyclic pairings (AB, BC, CD, AD) to be near the native values and varied the diagonal pairing values (AC, BD) to explore other packing options. We adjusted the area losses for our symmetrical model bundle protein to utilize the assumed sameness of all cyclic helix-helix pairing. We set the solvent-accessible area losses of pairings AB, BC, CD, and AD to 600 Å^2 (an approximate average of the values found from the hemerythin-myohemerythin fit (Yapa and Weaver, 1992, 1996)) and we set the non-native pairings AC and BD to have various area losses (see Table 2). We considered the AB, BC, CD, and AD pairings to be the native ones and therefore, state 16, with all four native pairings and no other pairings, to be the native state.

In simulation 1 (Fig. 4) we set the non-native pairings to have area losses of 285 Å^2. We also set to zero the forward rates from those kinetic states that would lead to a sterically hindered state, i.e., if pairing AC then not pairing BD, and vice versa. Fig. 4 shows that three states dominate the final fold at equilibrium. A system of noninteracting four-helix

TABLE 2 Area losses in Å^2 upon pairing

Pairings	Sim 1	Sim 2	Sim 3
AB	600	600	600
BC	600	600	600
CD	600	600	600
AD	600	600	600
BD	285	350	350
AC	285	350	285

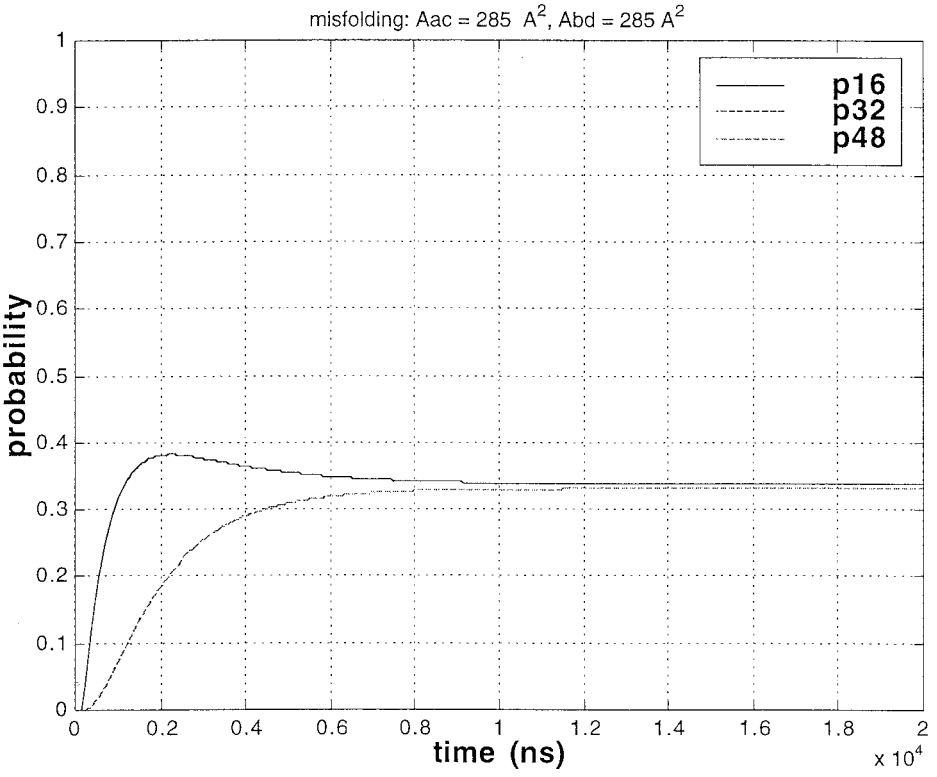

FIGURE 4 Probability of important kinetic intermediates versus time for a four-helix bundle example with a symmetrical pairing arrangement (see text). The area losses are 600 Å^2 for the four cyclic pairings (AB, BC, CD, and AD) and 285 Å^2 for the area losses of the AC and BD pairings. State 16, *solid line*; state 32, *dashed line*; state 48, *dashed-dotted line*.

bundle protein molecules will be divided equally among state 16 (native pairings), state 32 (native pairings + BD), and state 48 (native pairings + AC). Thus, the final fold is an equilibrium among the two five-pair states having the four native pairings and either the AC (state 48) or BD (state 32) pairing, and state 16, the native final state. State 16, with only native pairs, reaches a probability of ~0.38 in two μs. State 16 and states 32 and 48 attain equilibrium probability values of 0.33 on a 10 μs time scale. The simulation shows that if the AC or BD pairing has substantial buried area, it has stability and kinetic intermediates containing these pairings will persist and not unfold. If we consider the four-pair state to be the native structure, the protein has overfolded.

In simulation 2 (Fig. 5), the area loss in the non-native pairings was increased to 350 Å^2. Fig. 5 shows that states 32 and 48 dominate at equilibrium, each reaching a probability of 0.5 on a 10 μs time scale. The increase in the buried area of the AC and BD pairings to 350 Å^2 makes these pairings more likely to form and more likely to persist, compared to the area losses of 285 Å^2 in simulation 1. This results in overfolded final states and a minimally populated native state at equilibrium. The probability of state 16, the four-pair state containing only native pairings, peaks at ~0.22 after 1.6 μs and decays to almost zero at equilibrium.

In simulation 3 (Fig. 6), the two overfolded states 32 and 48 are given asymmetrical solvent-accessible area losses. The AC pairing is given an area loss of 285 Å^2 and the BD pairing is given an area loss of 350 Å^2. We see an initial increase in both states 32 (BD) and 48 (AC), but state 48, with the native and AC pairings, peaks at a probability of 0.23 at ~5 μs and slowly decays to almost zero probability, while state 32 (native + BD pairing) gains most of the probability. State 16, with only native pairs, peaks at a probability of 0.28 around 1.6 μs and slowly decays to almost zero probability. Because of the larger buried area, state 32 is more stable than state 48. In this case, one overfolded state is preferred over others. State 32 reaches 90% of the total probability at around 35 μs and has a probability of 0.92 at equilibrium. An identical result with the roles of states 32 and 48 reversed is found when the area losses for the AC and BD pairings are exchanged (results not shown).

In Fig. 7 the equilibrium probabilities at the three major equilibrium states, 16, 32, and 48, are plotted versus the solvent-accessible area loss given to each of the misfolded pairs (AC and BD). For small non-native area loss, state 16 dominates the equilibrium folding. As the non-native area loss increases, the overfolded states 32 and 48 increase in importance, reaching parity with state 16 at ~285 Å^2 and dominating the equilibrium probability thereafter. The strong change in behavior in the equilibrium values of the four- and five-pair states 16, 32, and 48 is directly related to

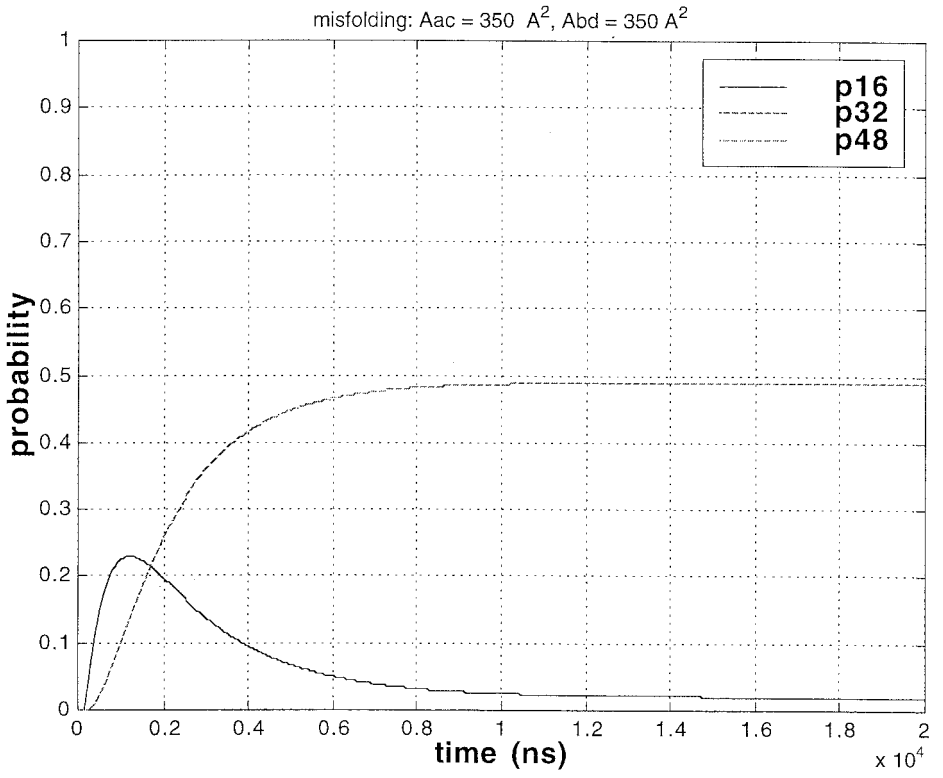

FIGURE 5 Probability of important kinetic intermediates versus time for a four-helix bundle example with a symmetrical pairing arrangement (see text). The area losses are 600 Å^2 for the four cyclic pairings (AB, BC, CD, and AD) and 350 Å^2 for the area losses of the AC and BD pairings. State 16, *solid line*; state 32, *dashed line*; state 48, *dashed-dotted line*.

the amount of buried hydrophobic area (area loss), with higher buried areas for the native pairings favoring folding and higher buried areas for the non-native pairings favoring overfolding.

DISCUSSION

Protein folding studies are becoming more detailed and recent experiments indicate that non-native kinetic intermediates occur in some cases (Ikai and Tanford, 1973; Khorasanizadeh et al., 1996; Baldwin, 1996). In a diffusion-collision model microdomain picture of folding, non-native collisions among microdomains are a natural consequence of the diffusive nature of the folding dynamics. We have studied the folding of four-helix bundle proteins with identical helices (the microdomains). This is a useful class of proteins in which to investigate non-native helix-helix pairings because four-helix bundle proteins are relatively simple, with a maximum of six helix-helix pairings and with many known protein structures in the protein data bank. In the diffusion-collision model, folding of the symmetrical four-helix bundle Regan-DeGrado (1988) protein proceeds by the successive pairing (coalescence) of the helices, the native pairings being AB, BC, CD, and AD, and the possible non-native pairings being AC and BD. The diffusion-collision model suggests two places in which folding kinetics could differ with nevertheless the same final native state, namely differences in the folding rates of kinetic intermediates due to differences in the intrinsic stabilities of microdomains and/or the lengths of loops between microdomains, and differences in the unfolding rates of kinetic intermediates due to differences in microdomain-microdomain packing as measured by solvent-accessible area loss. We find that there is a somewhat delicate balance between the importance of native and non-native helix-helix pairings, with the determining factor being the amount of solvent-accessible area loss upon pairing of two helices. Solvent-accessible area loss appears in the diffusion-collision model in both the folding and unfolding rates, with the latter being more sensitive due to exponential functionality (see Eq. 2 in Methods). We conclude that solvent-accessible area loss upon pairing of helices in a helical protein is an important determinant of what kinetic intermediates may occur in folding and what states persist at equilibrium. It appears that by mutating residues in a possible helix-helix interface that non-native pairings could occur in the final equilibrium structure of a four-helix bundle protein.

Fig. 3 shows the complete set of possible diffusion-collision states (as boxes) and transitions (as arrows between boxes) for a four-helix bundle protein composed of

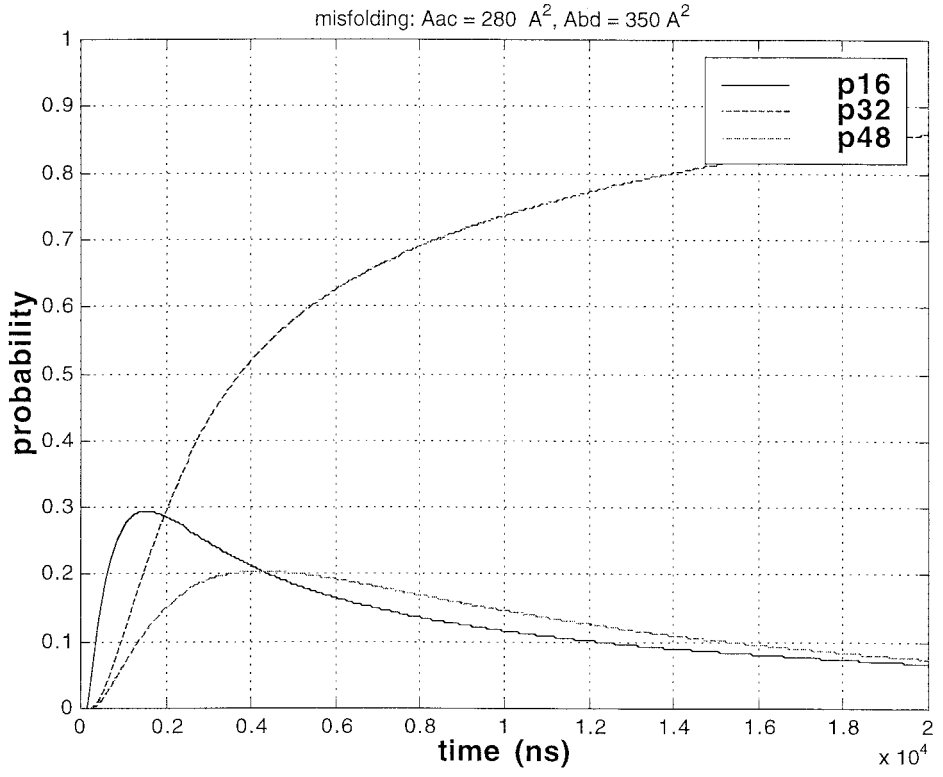

FIGURE 6 Probability of important kinetic intermediates versus time for a four-helix bundle example with a symmetrical pairing arrangement (see text). The area losses are 600 Å^2 for the four cyclic pairings (AB, BC, CD, and AD) and 350 Å^2 for the area loss of the BD pairing, 285 Å^2 for the area loss of the AC pairing. State 16, *solid line*; state 32, *dashed line*; state 48, *dashed-dotted line*.

helices A, B, C, and D and including all six possible pairings (AB, BC, CD, AD, AC, and BD). The most important diffusion-collision model folding states (as measured by maximum probability) and transitions for the three simulations are shown in Fig. 8. Compared with Fig. 3, in Fig. 8 only those states with a probability $p \geq 0.05$ at any time in the folding simulations are shown. In Fig. 8 the one-pair states are 3 (AD), 5 (CD), and 9 (AB). The two-pair states are 7 (BC–CD), 11 (AB–CD), and 13 (AB–BC). The three-pair states are 8 (BC–CD–AD), 12 (AB–CD–AD), 14 (AB–BC–CD), and 15 (AB–BC–CD). The four-pair states are 16, the state with all of the native pairs (AB–BC–CD–AD) and two states 31 with AD replaced by BD and 47 with AD replaced by AC. State 32 has the native set of pairs plus BD and state 48 has the native set of pairs plus AC. Both states 32 and 48 are five-pair states. The three states (16, 32, and 48) remain at equilibrium to varying degrees, depending on the folding conditions (see Fig. 4). In fact, in simulation 3, state 32 (the misfolded state with all native pairings and the non-native BD pairing) is the dominant species at equilibrium (a probability of 0.92).

Our study suggests that to avoid misfolding of the kind considered in this work, it may be possible to modify microdomain-microdomain pairings by altering the amino acid content of the microdomains. By identifying the important non-native intermediates, it may be possible to engineer the protein with mutations to reduce their importance (e.g., make their contacts unfavorable). In this way, non-native pairings may be designed to have weaker interactions than native pairings. The method could also be applied to increase yields in protein production in bulk, where intermolecular interactions between folding proteins cause disruption of the unimolecular folding process. The diffusion-collision model has already been successfully applied to bimolecular folding by Myers and Oas (1999), who applied the model to the dimerization of GCN4-p1 with success.

The diffusion-collision model emphasizes the collision and coalescence of microdomains in protein folding. Fig. 3 shows that in the model there are many ways of making the successive pairings of microdomains leading to the native state; Fig. 8 shows that some sequences of making pairings are more probable than others. Individual amino acid residues are deemphasized in the diffusion-collision model, except as they affect microdomain properties. By comparison, the "new view" of protein folding (Baldwin, 1994) emphasizes the free energy surface of the protein, with a general bias toward the native structure. It is useful to express diffusion-collision model reactions in energy terms. The initial stages of diffusion-collision folding are uphill in free energy due to the loss of volume (loss of entropy) in

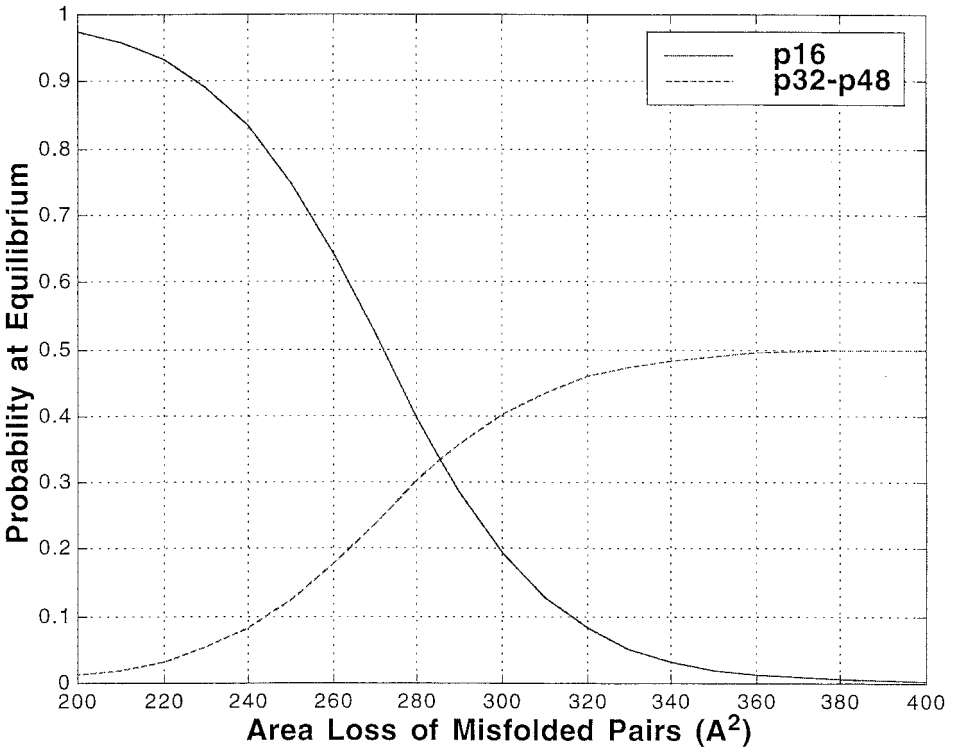

FIGURE 7 Equilibrium probabilities of the important kinetic intermediates versus solvent accessible area loss of the non-native pairs AC and BD. State 16, *solid line*; states 32-48, *dashed line*.

which two microdomains move around as they approach one another. At the same time, the microdomains are fluctuating in free energy as they move into and out of secondary structural conformations (helices in this paper). When two microdomains collide, there is an entropic barrier to be overcome due to the need for each microdomain to have its secondary structure somewhat correctly formed during a collision for coalescence to take place (determined by β). After surmounting the barrier, the folding protein falls into a free energy well (enthalpic) of depth determined by the attractive hydrophobic interaction between the microdomains. The folding protein then proceeds to the next pairing and so on until the final state is reached. Each diffusion-collision folding path between state 1 (no pairs) and the final state will be qualitative like the one described, but differing in 1) the order of pairing of microdomains, 2) the specific β-barrier to coalescence, and 3) the resulting hydrophobic interactions between specific microdomains. For a four-helix bundle protein with four helix-helix pairings, there will be 24 such paths between states 1 (no pairings) and 16 (four pairings); for five pairings there will be 120 paths between states 1 (no pairings) and 32 (five pairings);

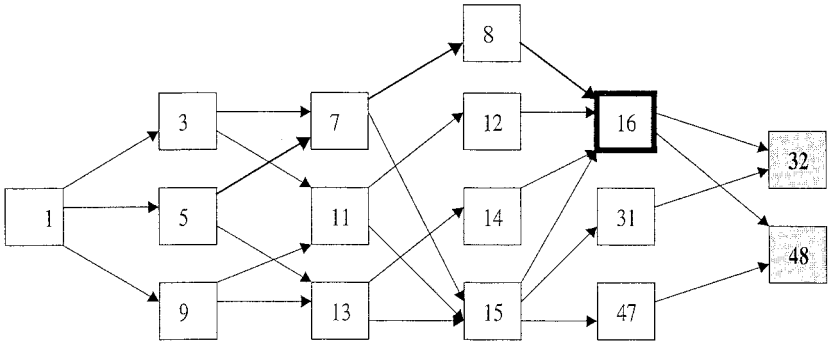

FIGURE 8 Diffusion-collision state diagram showing only those kinetic states that attain a probability of 0.05 or greater during the simulations. The various states are described in the text. The folding protein begins at state 1 (no pairs) and ends at equilibrium with probability in states 16 (the native state outlined with a dark border) and states 32 and 48, each with one non-native intermediate (shaded gray).

and for six pairings there will be 720 possible paths between states 1 (no pairings) and 64 (six pairings) (see Table 2).

The authors thank Dr. Rohit V. Pappu for discussions.

REFERENCES

Baldwin, R. L. 1994. Matching speed and stability. *Nature.* 369:183–184.

Baldwin, R. L. 1996. On-pathway versus off-pathway folding intermediates. *Fold. Des.* 1:R1–R8.

Bashford, D., F. E. Cohen, M. Karplus, I. D. Kuntz, and D. L. Weaver. 1988. Diffusion-collision model for the folding kinetics of myoglobin. *Proteins.* 4:211–227.

Bashford, D., D. L. Weaver, and M. Karplus. 1984. Diffusion-collision model for the folding kinetics of the λ-repressor operator binding domain. *J. Biomol. Struct. Dyn.* 1:1243–1255.

Burton, R. E., J. K. Meyers, and T. G. Oas. 1998. Protein folding dynamics: quantitative comparison between theory and experiment. *Biochemistry.* 37:5337–5343.

Capaldi, A. P., M. C. R. Shastry, C. Kleanthous, H. Roder, and S. E. Radford. 2001. Ultrarapid mixing experiments reveal that Im7 folds via an on-pathway intermediate. *Nature Struct. Biol.* 8:68–72.

Chothia, C. 1974. Hydrophobic bonding and accessible surface area in proteins. *Nature.* 248:338–339.

Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science.* 282:740–745.

Ferguson, N., A. P. Capaldi, R. James, C. Kleanhous, and S. E. Radford. 1999. Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* 266: 1597–1608.

Fezoui, Y., D. M. Hartley, D. M. Walsh, D. J. Selkoe, J. J. Osterhout, and D. B. Teplow. 2000. A de novo designed helix-turn-helix peptide forms nontoxic amyloid fibrils. *Nat. Struct. Biol.* 7:1095–1099.

Ikai, A., and C. Tanford. 1973. Kinetics of unfolding and refolding of proteins I. Mathematical analysis. *J. Mol. Biol.* 73:145–163.

Jackson, S. 1998. How do small single-domain proteins fold? Fold. Des. 3:R81–R91.

Karplus, M., and D. L. Weaver. 1976. Protein-folding dynamics. *Nature.* 260:404–406.

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

Karplus, M. and D. L. Weaver. 1994. The diffusion–collision model and experimental data. *Protein Sci.* 3:650–669.

Khorasanizadeh, S., I. D. Peters, and H. Roder. 1996. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nat. Struct. Biol.* 3:193–205.

Kragelund, B. B., P. Hojrup, M. S. Jensen, C. K. Schjerling, E. Juul, J. Knudsen, and F. M. Poulsen. 1996. Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.* 256:187–200.

Kragelund, B. B., P. Osmark, T. B. Neergaard, J. Schiodt, K. Kristiansen, J. Knudsen, and F. M. Poulsen. 1999. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat. Struct. Biol.* 6:594–601.

Kragelund, B. B., C. V. Robinson, J. Knudsen, C. M. Dobson, and F. M. Poulsen. 1995. Folding of a four-helix bundle: studies of acyl-coenzyme A binding protein. *Biochemistry.* 34:7217–7224.

Lacroix, E., A. R. Viguera, and L. Serrano. 1998. Elucidating the folding problem of α-helices: local motifs, long-range electrostatics, ionic strength dependence, and prediction of NMR parameters. *J. Mol. Biol.* 284:173–191.

Myers, J. K., and T. G. Oas. 1999. Reinterpretation of GCN4–p1 folding kinetics: partial helix formation precedes dimerization in coiled coil folding. *J. Mol. Biol.* 289:205–209.

Munoz, V., and L. Serrano. 1994a. Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* 1:399–409.

Munoz, V., and L. Serrano. 1994b. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J. Mol. Biol.* 245:275–296.

Munoz, V., and L. Serrano. 1994c. Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J. Mol. Biol.* 245:297–308.

Munoz, V., and L. Serrano. 1997. Development of the multiple sequence approximation within the AGADIR model of α-helix formation. Comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers.* 41:495–509.

Paliakasis, C. D., and M. Kokkinidis. 1992. Relationships between sequence and structure for the four-α-helix bundle tertiary motif in proteins. *Protein Eng.* 5:739–748.

Pappu, R. V., and D. L. Weaver. 1998. The early folding kinetics of apomyoglobin. *Protein Sci.* 7:480–490.

Regan, L., and W. F. DeGrado. 1988. Characterization of a helical protein designed from first principles. *Science.* 241:976–978.

Rojnuckarin, A., S. Kim, and S. Subramaniam. 1998. Brownian dynamics simulations of protein folding: access to milliseconds time scale and beyond. *Proc. Natl. Acad. Sci. U.S.A.* 95:4288–4292.

Sheridan, R. P., R. M. Levy, and F. R. Salemme. 1982. α-Helix dipole model and electrostatic stabilization of 4-α-helical proteins. *Proc. Natl. Acad. Sci. U.S.A.* 79:4545–4549.

Vasilkoski, Z., and D. L. Weaver. 2000. A generator of protein folding kinetic states for the diffusion-collision model. *J. Comp. Chem.* 21: 923–932.

Wagner, C., and T. Kiefhaber. 1999. Intermediates can accelerate protein folding. *Proc. Natl. Acad. Sci. U.S.A.* 96:6716–6721.

Wallis, R., A. Reilly, A. Rowe, G. Moore, R. James, and C. Kleanthous. 1992. In vivo and in vitro characterization of overproduced colicin E9 immunity protein. *Eur. J. Biochem.* 207:687–695.

Weber, P. C., and F. R. Salemme. 1980. Structural and functional diversity in 4-α-helical proteins. *Nature.* 287:82–84.

Yapa, K., and D. L. Weaver. 1992. Folding kinetics of designer proteins: application of the diffusion-collision model to a de novo designed four-helix bundle. *Biophys. J.* 63:296–299.

Yapa, K., and D. L. Weaver. 1996. Protein folding dynamics: application of the diffusion collision model to the folding of a four-helix bundle. *J. Phys. Chem.* 100:2498–2509.