Biochemistry

© Copyright 1998 by the American Chemical Society Volume 37, Number 16 April 21, 1998

Accelerated Publications

Protein Folding Dynamics: Quantitative Comparison between Theory and Experiment†

Randall E. Burton, Jeffrey K. Myers, and Terrence G. Oas*

Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710

Received January 30, 1998; Revised Manuscript Received March 6, 1998

ABSTRACT: The development of a quantitative kinetic scheme is a central goal in mechanistic studies of biological phenomena. For fast-folding proteins, which lack experimentally observable kinetic intermediates, a quantitative kinetic scheme describing the order and rate of events during folding has yet to be developed. In the present study, the folding mechanism of monomeric λ repressor is described using the diffusion-collision model and estimates of intrinsic α-helix propensities. The model accurately predicts the folding rates of the wild-type protein and five of eight previously studied Ala ↔ Gly variants and suggests that the folding mechanism is distributed among multiple pathways that are highly sensitive to the amino acid sequence. For example, the model predicts that the wild-type protein folds through a small number of pathways with a folding time of 260 μs. However, the folding of a variant (G46A/G48A) is predicted to fold through a large number of pathways with a folding time of 12 μs. Both folding times quantitatively agree with the experimental values at 37 °C extrapolated to 0 M denaturant. The quantitative nature of the diffusion-collision model allows for rigorous experimental tests of the theory.

In recent years, many experimental and theoretical studies have focused on the problem of describing the mechanism of protein folding. An important and elusive goal is the development of folding models that make quantitative predictions of protein folding and unfolding rates as a function of amino acid sequence, buffer composition, and temperature. Such a model requires detailed understanding of the factors that determine folding and unfolding rates. A model with this potential is the diffusion-collision model of Weaver and Karplus (*1*, *2*). Diffusion-collision theory breaks a protein down into elements of secondary structure (microdomains) that diffuse through solution and collide to form progressively larger microdomains until the folding reaction is complete. Because microdomains can form and coalesce in many possible combinations, there are many possible paths. This description is consistent with the "new view" of folding, which emphasizes the multiplicity of folding routes rather than a single pathway followed by all molecules (*3*, *4*). Indeed, the model keeps track of a large number of possible intermediates in and can calculate their relative populations at any point in time, starting from a single populated state.

A salient feature of the diffusion-collision model is the parameter β, the probability that a collision event between two microdomains will lead to productive folding. β is related to the amount of native structure present in the isolated microdomain, determined by its intrinsic secondary structure content. One limitation of the model is the lack of experimental measurements of the intrinsic stability of the isolated elements of secondary structure. Previous diffusion-collision-based calculations of N-terminal domain of λ repressor (*5*) and myoglobin (*6*) folding kinetics lacked estimates of actual stabilities of individual microdomains (in the case of these two proteins, α-helices). Instead, calcula-

† This work was supported by the National Institutes of Health (Grant GM45322 to T.G.O. and the postdoctoral fellowship F32 GM18957 to J.K.M.).

* Corresponding author e-mail: oas@biochem.duke.edu.

S0006-2960(98)00245-1 CCC: $15.00 © 1998 American Chemical Society
Published on Web 04/03/1998

tions were performed using several hypothetical values of β, which was assumed to be the same for all helices. Clearly, individual helices within a protein have different intrinsic stabilities, making this assumption unrealistic. Realistic calculations of rates require reasonable estimates of the intrinsic stability of each helix.

One approach to measure individual β values would be to make peptides models of each helix and measure the helical content using CD or NMR (*7*). However, peptides taken from helices in proteins often show little helical structure, making accurate determination of fractional helicity difficult. In addition peptide fragments are prone to aggregation. Fortunately, it is possible to calculate helix propensities using the algorithm AGADIR (*8*−*11*), which is based on experimental studies of helical peptides. AGADIR combines two long-standing lines of research: (1) helix/coil transition theory and (2) experimental studies of helix formation in peptides (*12, 13*). Various parameters representing the factors contributing to helix stability (intrinsic propensities, side-chain interactions, etc.) have been calculated by fitting experimental helix contents to helix/coil theory. AGADIR predicts the helical content of previously studied peptides to an accuracy of ±12% (*9*). This gives a promising means of calculating β from the predicted helix content of the isolated α-helices. However, there are many interactions potentially involved in helical stability. Specific interactions not included in the algorithm that contribute to the helical stability of a particular sequence could lead to underprediction of the helix content of that sequence. Additional problems may be caused by the very low helix content of some of the λ repressor helices (see below). Therefore, an experimental measure of the intrinsic stability of the helices is still preferable, and the AGADIR predictions should be considered as estimates. Work is currently underway to develop an experimental model of monomeric λ repressor in which the tertiary interactions are removed without affecting the intrinsic stabilities of the helices. We hope to use NMR studies on these peptides to confirm the AGADIR predictions.

In this paper, the diffusion-collision model has been applied to λ_{6-85}, a monomeric version of the N-terminal domain of λ repressor. This protein folds on a very fast (submillisecond) time scale (*14, 15*). Helical contents were calculated by AGADIR for each helix in the G46A/G48A variant and have been used to determine β values in the diffusion-collision model. These β values are used to calculate the formation rates of progressively more helical conformations. In combination with estimated breakdown rates, overall time dependence of native structure formation is calculated and compared with experimentally measured values for the wild-type protein and eight variants whose folding kinetics have been previously determined (*16*).

MATERIALS AND METHODS

In our implementation of diffusion-collision theory, the five helical regions of λ_{6-85} identified by NMR (*17*) are treated as individual microdomains, along with any interacting combination of these helices, which represent multihelical microdomains. Partially folded conformations in the folding process are represented as species with various combinations of α-helices formed and bonded together by nativelike, tertiary interactions. State 1 is the unfolded state, which contains all completely unfolded conformations, as well as those with a single helix formed. The latter are included because isolated helices are only marginally stable and intercovert with coil conformations on a nanosecond time scale. There are 51 additional states that correspond to various combinations of native helix−helix interactions.

Interconversion between these states occurs through coalescence of microdomains to form higher-order conformations and the breakdown of microdomains into smaller units. The rate of productive collision (coalescence) of two microdomains to form a bond, $1/\tau_f$, is calculated using the following equation from diffusion-collision theory (*2*):

$$\tau_f = \frac{G}{D} + \frac{LV(1-\beta)}{DA\beta} \tag{1}$$

$$G = \frac{R_{max}\left(1 - \frac{9}{5}\epsilon + \epsilon^2 - \frac{1}{5}\epsilon^3\right)}{3\epsilon(1-\epsilon^3)} \qquad \epsilon = \frac{R_{min}}{R_{max}}$$

$$\frac{1}{L} = \frac{1}{R_{min}} + \alpha\left(\frac{\alpha R_{max} \times \tanh[\alpha(R_{max} - R_{min})] - 1}{\alpha R_{max} - \tanh[\alpha(R_{max} - R_{min})]}\right)$$

$$\alpha = \frac{1}{\sqrt{D\tau_C}}$$

$$V = \frac{4}{3}\pi(R_{max}^3 - R_{min}^3)$$

$$D = \frac{k_B T}{6\pi\eta}\left(\frac{1}{R_a} + \frac{1}{R_b}\right)$$

$$R_{min} = R_a + R_b \qquad R_{max} = R_{min} + \text{linker length}$$

where D is the relative diffusion coefficient, G and L are sequence-invariant parameters that satisfy boundary conditions for the diffusion equation in three dimensions. The diffusing microdomain is modeled as the smallest sphere that can contain all of the atoms with a packing ratio of 0.7 (*5*). A is the sum of the areas of these two spheres, and R_a and R_b are the radii. τ_C is the time constant for helix ↔ coil transitions, which has been determined experimentally to be approximately 10^{-8} s (*18*−*20*), and β is the probability that both microdomains have enough nascent structure to collide productively. This probability is calculated from the product of the individual formation probabilities of the two colliding microdomains. In our implementation, the formation probabilities (F_{60}) of the single-helix microdomains are calculated from standard Lifson-Roig helix−coil theory, assuming that any helical segment at least 60% of the length of the native helix would lead to a productive collision. This calculation is performed using the homopolymer approximation (*21*) and average helicities predicted by AGADIR-2s for the five helical segments, as listed in Table 1 for the G46A/G48A variant. The effect of Ala → Gly substitutions is calculated using single-site perturbative Lifson-Roig theory (*21*) and an w_{Ala}/w_{Gly} ratio of 35 (*22*). For eight such variants (including the A46G/A48G wild-type sequence), Table 2 lists the altered F_{60} values obtained for the helices at the site of substitution. In species with multihelical microdomains, the formation probability is assumed to be 1, reflecting the

Table 1. Parameters Used in the Diffusion-Collision Model of the Folding of λ_{6-85}[a]

helix	residues	% helix	w	F_{60}	linker length to $(n+1)$
1	9−30	41	1.26	0.35	2
2	33−39	5.3	1.50	0.04	4
3	44−52	3.1	1.10	0.03	6
4	59−69	49	1.85	0.43	3
5	79−85	0.89	0.84	0.01	

[a] The boundaries of the five helical regions, shown in the first column, differ slightly from those reported in the crystal structure (*25*). The added residues are in a helical conformation in the native state, but do not fit the most conservative definitions of α-helix. The overall helix propensities of these regions were calculated using AGADIR (*8*). Sequences corresponding to the individual helices in λ_{6-85} were submitted to the program, with pH set to 7.0, ionic strength to 0.1 M, and temperature to 310 K. The % helicity was then averaged over the residues in the helix. Using Lifson-Roig theory, a uniform w value was calculated which predicted the same % helix as AGADIR, with the nucleation parameter $v^2 = 0.003$. With these values, it was possible to calculate the probability that 60% of the residues are in a helical conformation (F_{60}) for each helix. The last column is the number of residues from the end of a given helix to the beginning of the next. These are determined from the boundaries listed in the first column, except for the linker between helices 4 and 5. This stretch of 10 residues is not helical, but clearly contains nonextended structure in the C-terminal seven residues of the linker, including a proline in the N-cap position of helix 5. To account for the potential lack of flexibility in this linker, the length of the 4−5 linker was shortened from 10 to 3 residues.

Table 2: Comparison of Folding Times Predicted by the Diffusion-Collision Model and Those Observed Experimentally[a]

variant	helix changed	new F_{60}	predicted τ_f ($\times 10^6$ s)	experimental τ_f ($\times 10^6$ s)
G46A/G48A			12	12 ± 2
M15	1	0.042	25	100 ± 13
M20	1	0.033	27	17 ± 2
M37	2	0.0007	21	10 ± 3
M49	3	0.005	25	17 ± 1
M63	4	0.026	25	18 ± 2
M66	4	0.061	21	190 ± 40
M81	5	0.0005	110	16 ± 4
WT	3	0.0002	260	204 ± 25

[a] The "M#" variants are single Ala → Gly substitutions in the mutant (G46A/G48A) background, where # is the site of the substitution. For each variant, the helix which contains the substitution is listed, along with the new F_{60} value corresponding to the mutant sequence. The experimental folding times are taken from previous NMR spectra (*16*), which were reanalyzed using a more sophisticated lineshape simulation package (*41*) and extrapolated to 0 M urea. The initial values for the folding calculation are such that only the completely unfolded state is populated. The predicted folding time (τ_f) is the time at which the relative population of the native state reaches 0.6, as described previously (*5*).

cooperative stabilization of structure resulting from tertiary interactions. The β value for a pairwise collision is the product of F_{60} values ($\beta_{ij} = F_{60,i}F_{60,j}$).

The dissociation rates of the helix−helix bonds are calculated using the transition-state approximation:

$$\frac{1}{\tau_u} = k_u = \nu e^{-\Delta G_{int}/k_B T} \qquad (2)$$

where ν is the rate of dissociation in the absence of an energy barrier, k_B is the Boltzmann constant, T is the absolute temperature, and ΔG_{int} is the interaction energy between the helices that are involved in the bond. This energy is calculated using the structural thermodynamics method of Freire and co-workers (*23*, *24*). The polar and apolar surface areas buried by helix−helix contacts were determined from the crystal structure (*25*) using the algorithm of Richmond and Richards (*26*), incorporated in the ACCESS program (S. Presnell). ν is set to 10^{10} s^{-1}, consistent with observed rates of bimolecular dissociation without an energy barrier (*27*). Determining an appropriate value for ν will be essential to describe the folding mechanism in detail, as this parameter uniformly affects all microdomain breakdown rates which then determine the extent to which intermediate states accumulate during the reaction. Table 3 lists parameters used to calculate association and dissociation rate constants for typical microdomains in λ_{6-85}.

RESULTS

The rates of microdomain coalescence are calculated using these parameters and eq 1. The resulting 52 × 52 rate matrix represents a linear system of first-order differential equations which describe the evolution of the system with time. The time-dependent solution of these differential equations is found using a numerical matrix method. The solution for each state appears as a sum of exponential terms, whose amplitudes are proportional to the eigenvectors of the rate matrix and time constants are the corresponding eigenvalues (*5*).

These calculated populations for the wild-type and G46A/G48A sequences are plotted versus time in Figure 1. The model predicts that very few partially folded species are ever significantly populated. In wild-type, the substate with all but helix 3 formed is predicted to become transiently populated at about 50 μs. In the G46A/G48A variant, several species are predicted to exist transiently in the first 10 μs of folding. Other than these species, the only other substates of either protein that are significantly populated over the course of the folding reaction are the nonhelical starting substate and the fully native species. None of the intermediate states are significantly populated at equilibrium in this model, thus making them invisible to the NMR techniques used to measure the folding rates of these proteins (*15*). Even the wild-type kinetic intermediate, which is predicted to be nearly 70% populated during the reaction, is predicted to have an equilibrium population of less than 1%, too low to significantly affect the NMR analysis. This observation supports Zwanzig's recent analysis that equilibrium two-state behavior does not preclude conformational or kinetic complexity (*28*).

The overall refolding time constant is taken as the length of required time to attain a native-state population of 0.6. As shown in Table 2, the model described above accurately predicts the folding times of the wild-type protein and five variants by only including the effects of these sequence changes on the intrinsic helix propensities. The two variants whose experimentally observed folding times are significantly longer that the current model predicts (A15G and A66G) involve substitutions at completely buried residues in the protein. These substitutions also destabilize the protein by more than the ~0.8 kcal/mol observed for the surface Ala → Gly substitutions (Daugherty, et al., in preparation). The longer folding times for these variant might indicate that tertiary interactions affect the collision rates of microdomains.

Table 3: Parameters Used to Calculate Forward (Association) and Reverse (Dissociation) Rate Constants for Selected Microdomains in λ_{6-85}[a]

reaction	1 + 2 ⇔ 12	1 + 45 ⇔ 145	12 + 45 ⇔ 1245	123 + 45 ⇔ 12345 (N)
R_1 (Å)	9.63	9.63	10.5	11.3
R_2 (Å)	6.38	9.37	9.37	9.37
R_{min} (Å)	16.0	19.0	19.8	20.7
R_{max} (Å)	23.0	117.	86.3	41.7
D (Å^2/s)	8.53×10^{10}	6.89×10^{10}	6.62×10^{10}	6.38×10^{10}
G (Å^2)	6.29×10^{1}	2.07×10^{4}	6.97×10^{3}	4.40×10^{2}
L (Å)	83.4	11.0	11.3	18.7
ΔV (Å^3)	3.38×10^{4}	6.68×10^{6}	2.66×10^{6}	2.67×10^{5}
A (Å^2)	1.68×10^{3}	2.27×10^{3}	2.49×10^{3}	2.71×10^{3}
β	0.0157	0.352	1	1
k_f (s^{-1})	9.4×10^{4}	3.8×10^{5}	9.5×10^{6}	1.4×10^{8}
ΔG_{int}	1.9	12.	15.	16.
k_u (s^{-1})	6.2×10^{8}	8.4×10^{1}	4.5×10^{-1}	3.7×10^{-2}

[a] These parameters are determined from crystal structure data on the DNA-bound dimer (*25*) using previously described methods (*5*).

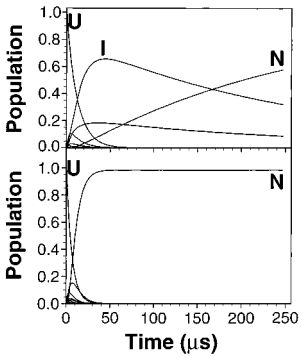

FIGURE 1: The solution of the system of diffusion equations describing all 52 substates determines the concentration of each state as a function of time. The predicted evolution of the system with time is shown for the wild-type sequence (top panel) and the G46A/G48A variant (bottom panel). The traces indicate the transient formation of intermediate states during the folding process. For the wild-type sequence, a kinetic intermediate (I) is predicted to build up significantly during the folding process. This intermediate contains all of the elements of the native structure except helix 3. When helix 3 is stabilized by the G46A/G48A substitution (lower panel), the intermediate disappears, and the overall folding rate is faster.

More likely, the removal of the Cβ methyl group disrupts the interface between microdomains and increases the breakdown rate. Modeling the tertiary effects of these substitutions presumably requires a more sophisticated treatment of interhelical interactions than provided by the current model. The effect of one substitution (A81G) is considerably overestimated by the model, suggesting that the helix−coil equilibrium is either not appropriately modeled or not important for the assembly of this region of the molecule. The loop from helix 4 to helix 5 is the longest in the protein and contains nonhelical local structures that would not be affected by a glycine residue at position 81. If this loop plays the critical role docking the C-terminal region of the protein onto the rest of the structure, the mutation would have subtle kinetic effects. This argument suggests that further improvements to the model should include the propensity of some loop sequences to form other local structures such as β-turns.

The reaction networks calculated using the diffusion-collision model for the wild-type and G46A/G48A sequences are shown in Figure 2. For the wild-type sequence, the model predicts that there are few efficient pathways from the unfolded state to the native state, shown as thick bars in Figure 2. In contrast, the G46A/G48A protein is predicted to have a wide variety of efficient pathways to chose from. The reason for this difference is the stabilization of helix 3 by the double-alanine substitution. In the wild-type case, productive collisions with helix 3 are rare, which dramatically reduces the rate constants for those collision. In the G46A/G48A case, the relative intrinsic stabilities of the five helices are less varied, and the order of collision events becomes more random.

There are several assumptions inherent to the current model that will be tested in future experimental and theoretical work. First, only native contacts are considered. Clearly, provisions must be made for misfolded structures, which would result in off-pathway kinetic "traps" (*29*). Second, the intrinsic helicities of the five α-helices in λ_{6-85} are inferred from the AGADIR program, which is unlikely to produce highly precise estimates for sequences with very low (<5%) intrinsic helicity. The model is very sensitive to the intrinsic stability of the least-helical segments. Third, the effects of Ala → Gly substitutions are modeled as single-site perturbations in an otherwise homogeneous peptide. It is unlikely that the intrinsic helix propensity of a sequence is evenly distributed throughout the peptide, lending a context dependence on the effect of a Ala → Gly substitution. Fourth, the linker regions between α-helices are modeled as completely flexible "strings" which allow random sampling of the diffusion volume. In reality, there may be local conformational preferences in the linker regions that favor microdomain collisions in the native orientation. Fifth, the microdomain breakdown rates are very crudely modeled. Breakdown rates are not critical under the strongly folding conditions used here, but experimental rate measurements are made in the transition region between folding and unfolding conditions. To accurately model folding rates under these conditions, it will be necessary to find reasonable breakdown rates for the microdomain−microdomain bonds. In particular, the concentrations of intermediate states are extremely dependent on the breakdown rates, so the model cannot accurately predict the maximum concentration of intermediate states. Despite these assumptions, the impressive correspondence between experimentally observed folding times and the diffusion-collision calculations demonstrate that the fundamental theory is sound.

DISCUSSION

Predicted Effects of Solvent Conditions. Because of the prominent role played by diffusion in this model, it is

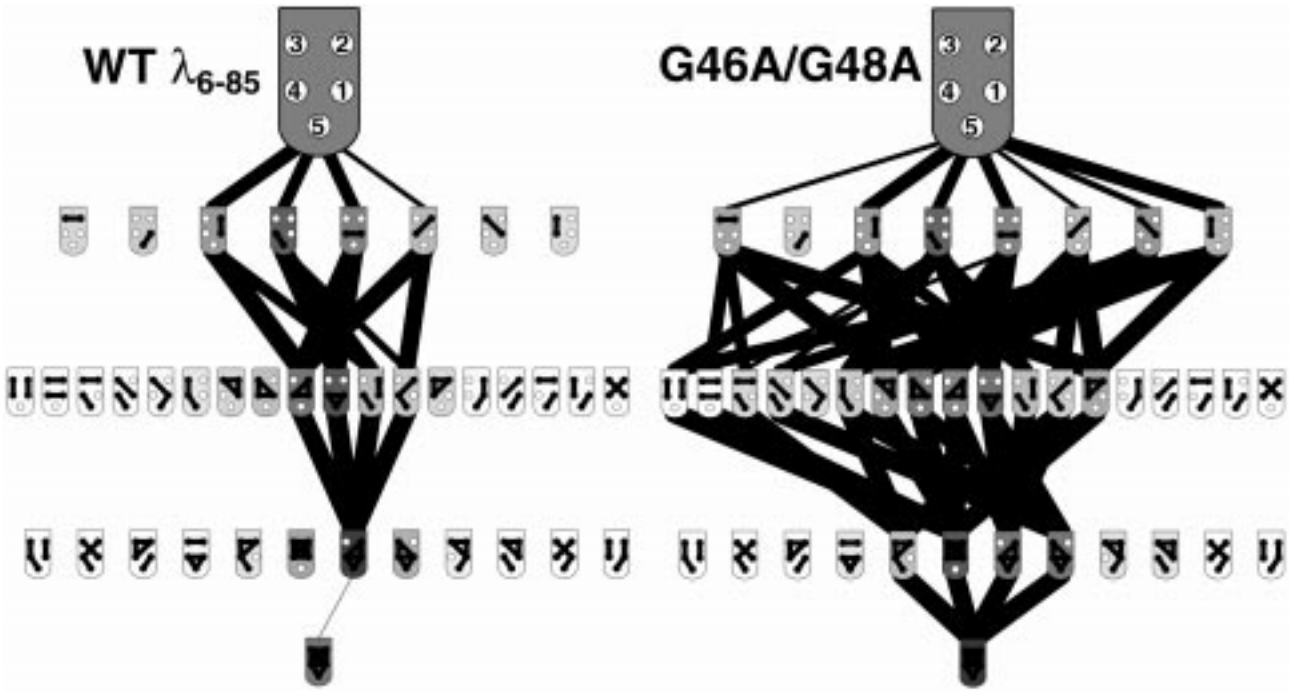

FIGURE 2: A schematic depiction of the results of diffusion-collision calculations for the wild-type λ_{6-85} sequence and the G46A/G48A variant, showing a subset of the 52 substates which are involved in kinetically relevant diffusion-collision events. The open and filled circles in the substate symbols represent nonhelical or helical segments, respectively. The bars between circles represent interactions between helical regions. The darkness of each substate symbol is proportional to the maximum concentration calculated for that species during the folding process. The thickness of the lines connecting the symbols is proportional to the rate constant for that transition. To emphasize the predominant pathways for each case, a pruning algorithm was applied that reduced the number of lines in the lower regions of the figures. Starting from the top (unfolded) state, the eight rate constants to the second layer were calculated. Substates whose formation rate constants were less than 1% of the maximal rate constant at that level were presumed to be kinetically unimportant and were not included in calculating rates to the next lower layer. The wild-type protein appears to fold via a few possible routes, all of which go through a common intermediate lacking helix 3, whereas the G46A/G48A variant is predicted to use a wide variety of folding pathways with no common intermediates.

possible to make some qualitative predictions about the effects of bulk solvent conditions such as temperature and viscosity. Temperature is predicted to have two opposing effects on the folding rate. First, folding should speed up because the diffusion constant will increase with temperature. However, the β parameter will go down as the helices are destabilized, which should slow folding. These competing effects may lead to a shallow and/or nonlinear temperature dependence of folding rate, which has been observed for some proteins (*30*, *31*), including the G46A/G48A variant of λ_{6-85} (unpublished observations). Another perturbant of the diffusion rate, solvent viscosity, should have the strongest effect on collisions between very stable microdomains, where β approaches 1. Inspection of eq 1 shows that under these circumstances τ_f simplifies to G/D. However, for collisions involving unstable helices, where β is small, the effect of viscosity is predicted to be less dramatic (*5*). As discussed below, collisions involving unstable helices are predicted to be the rate-limiting steps in the overall folding of λ_{6-85}, indicating that the overall folding rate should be insensitive to solvent viscosity.

Transition States and Diffusion-Influenced Reactions. The concept of a "transition state" for the rate-limiting step in protein folding has been used to explain kinetic folding and unfolding data for nearly three decades (*32*). However, it is not clear how to define a transition state in our current model. First, there is no single "pathway" for the protein to follow, especially for the G46A/G48A variant (Figure 2), which means there is not necessarily a unique rate-limiting step. Second, the "barrier" for the forward steps is the time required for two microdomains to "find" each other via Brownian motion in a viscous medium. As modeled, the barrier for the breakdown steps is the unfavorable solvation of the hydrophobic groups that constitute the helix−helix interface. Therefore, the rate-limiting steps for bond formation and breakage would appear to be different. However, a correspondence can be established by relating the "diffusive" barrier of the forward reaction to the loss of configurational entropy upon bringing two microdomains together and locking them into place. In this scenario, the rate-limiting step in both directions occurs as the microdomains are juxtaposed in the native orientation, but the favorable interface contacts have not yet been formed.

Role of Intermediates in Protein Folding. There has been considerable debate about the efficiency of folding with or without intermediates. Some have argued that stable intermediates inherently slow folding (*33*, *34*); others argue that they are essential for productive folding on a reasonable time scale (*35*). The diffusion-collision predictions shown in Figure 2 suggest that in λ_{6-85} the dominant intermediates formed precede the most difficult folding steps. For wild-type λ_{6-85}, the docking of helix 3 onto the rest of the structure is predicted to be rate limiting. Therefore, there is a significant buildup of an intermediate with all of the bonds formed between helices 1, 2, 4, and 5. When helix 3 is stabilized by the G46A/G48A substitutions, the model predicts that this intermediate disappears and the folding rate increases 20-fold. In the context of the productive/unproductive intermediate debate, this would be evidence for intermediates that slow folding. However, the predicted

folding mechanism for the G46A/G48A variant is widely distributed among the possible substates, generating a much larger number of intermediate states. It could be argued, therefore, that the double-alanine substitution has merely replaced one essential intermediate with a set of more productive intermediate substates.

The diffusion-collision model offers a new perspective on this debate. The essential argument for the productive/unproductive intermediate debate is whether the intermediate is on or off the folding pathway (*36*). The diffusion-collision model does not impose any order to the process; there are as many potential pathways as there are microscopic substates. Given this inherent complexity of the model, it is interesting that analysis of the folding calculation seems to indicate predominant pathways, whose number depends sensitively on amino acid sequence. The folding of λ_{6-85} tends to be dominated by the fastest possible route to the native state, but there are a number of possible pathways that are predicted to be close enough in energy that small sequence changes can force a switch in pathway. Any observed on-pathway intermediates simply reflect the easiest path to the native state, not necessarily the only one. Under some conditions, the diffusion-collision model predicts nearly two-state kinetic transitions for folding in which partially folded states are never populated to a significant extent (*5*), indicating that intermediates are not inherent to the model.

Funnels vs Pathways. The diffusion-collision model fits well with the "new view" of protein folding, dominated by funnel-shaped "energy landscapes" (*3, 4*). This view posits that an unfolded chain finds the native state because there is a general bias in the energy landscape favoring the formation of native contacts. In the diffusion-collision model, this bias is incorporated in the rapid formation of nativelike local contacts within a microdomain, which are then "frozen" in place by productive collisions, which provide stabilization through tertiary interactions. The "bottleneck" of the simple funnels proposed to represent the energy landscape of λ_{6-85} corresponds to a loss of configurational entropy without a concomitant increase in favorable contacts (*16*). In the diffusion-collision picture, this is represented as the formation of a marginally stable helix, just prior to its adhesion to the rest of the protein, restricting the entropy of intervening loop sequences.

Comparison with "Hydrophobic Collapse" Models. Several recent models of protein folding have proposed that the earliest step in folding is the contraction of the polypeptide chain due to the hydrophobic effect (*37, 35*). The formation of secondary structure is thought to occur either during or after this collapse event. The diffusion-collision model, with its "helix first" approach, is diametrically opposed to this kind of model. However, the distinction between hydrophobic collapse and helix formation is not so clear, since the formation of an α-helix buries a significant amount of apolar surface area (*38*). In fact, local hydrophobic interactions may be an important mechanism in helix nucleation. Thus, helices may represent locally collapsed microdomains, but hydrophobic collapse to produce nativelike, long-range tertiary interaction prior to the formation of secondary structure is not envisioned in the diffusion-collision description. A hydrophobic collapse model could be thought of as an extension of the diffusion-collision model which allows unfolded microdomains to collide and proceed to fold together without first diffusing apart. Such an extension will be necessary to extend the model to proteins with weak local structural preferences such as plastocyanin, a predominately β-sheet protein with few strong local conformational preferences (*39*).

Implications for Other Proteins. The N-terminal domain of λ repressor appears to be an ideal candidate for a diffusion-collision description. This model may be useful for other fast-folding helical proteins. Whether the model can be extended to include mixed α/β or all-β proteins is not clear. Less is known about the intrinsic stability of β structures than α-helices. Additionally, an α-helix involves local interactions between residues close in sequence. β-Sheets can come from strands located distant in sequence. However, the formation of β-sheets can also be influenced by local conformational preferences (*40*). Another element of structure poorly described by the model is the β-turn. The presence of an independently stable turn between microdomains would invalidate the assumption that the linker is a freely jointed chain and therefore would limit the diffusion space of the adjoining microdomains. Therefore even a turn or loop conformation which has weak intrinsic stability might restrict diffusion space enough to be important. Despite these potential limitations, the success of the model in predicting folding rates of λ_{6-85} and its variants should encourage attempts to apply the model to other types of proteins.

Summary. When combined with reasonable estimates of intrinsic helicity, the diffusion-collision model does a remarkable job of predicting both the absolute rates of folding for λ_{6-85} and the relative effects of Ala → Gly substitutions. To our knowledge, this is the first attempt to calculate the rates for a complete folding reaction of a protein and a series of variants. The success of these calculations largely reflects the progress made in understanding both helix−coil transitions and the energetics of protein folding. The strength of helix−coil theory makes λ_{6-85} an ideal candidate for studying diffusion-collision experimentally. Further refinements of the model will address the breakdown rates of microdomains as well as the effects of environmental changes (e.g., temperature, viscosity, chemical denaturants). *Mathematica* notebooks containing all of the parameters and functions used in these calculations are available from the authors.

ACKNOWLEDGMENT

The authors thank Luis Serrano for his help with AGADIR predictions for λ_{6-85} and David Weaver for his assistance with our implementation of diffusion-collision theory. We also thank David Baker, Robert Baldwin, Osman Bilsel, Chris Bystroff, Sina Ghaemmhagami, Bob Matthews, Kevin Plaxco, Marty Scholtz, and Jill Zitzewitz for their helpful comments on the manuscript prior to publication.

REFERENCES

1. Karplus, M., and Weaver, D. L. (1976) *Nature (London) 260*, 404−406.
2. Karplus, M., and Weaver, D. L. (1994) *Protein Sci. 3*, 650−668.
3. Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995) *Proteins 21*, 167−195.
4. Dill, K. A., and Chan, H. S. (1997) *Nat. Struct. Biol. 4*, 10−19.
5. Bashford, D., Weaver, D. L., and Karplus, M. (1984) *J. Biomol. Struct. Dynam. 1*, 1243−1255.

6. Bashford, D., Cohen, F. E., Karplus, M., Kuntz, I. D., and Weaver, D. L. (1988) *Proteins 4*, 211−227.
7. Reymond, R. T., Merutka, G., Dyson, H. J., and Wright, P. E. (1997) *Protein Sci. 6*, 706−716.
8. Muñoz, V., and Serrano, L. (1994) *Nat. Struct. Biol. 1*, 399−409.
9. Muñoz, V., and Serrano, L. (1995) *J. Mol. Biol. 245*, 275−296.
10. Muñoz, V., and Serrano, L. (1995) *J. Mol. Biol. 245*, 297−308.
11. Muñoz, V., and Serrano, L. (1997) *Biopolymers 41*, 495−509.
12. Scholtz, J. M., and Baldwin, R. L. (1992) *Annu. Rev. Biophys. Biomol. Struct. 21*, 95−118.
13. Kallenbach, N. R., Lyu, P., and Zhou, H. (1996) *Circular dicroism and the conformational analysis of biomolecules*; New York, Plenum Press. 202−259.
14. Huang, G. S., and Oas, T. G. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 6878−6882.
15. Burton, R. E., Huang, G. S., Daugherty, M. A., Fullbright, P. W., and Oas, T. G. (1996) *J. Mol. Biol. 263*, 311−322.
16. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L., and Oas, T. G. (1997) *Nat. Struct. Biol. 4*, 305−310.
17. Huang, G. S., and Oas, T. G. (1995) *Biochemistry 34*, 3884−3892.
18. Hammes, G. G., and Roberts, P. B. (1969) *J. Am. Chem. Soc. 91*, 1812−1816.
19. Williams, S., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H., and Dyer, R. B. (1996) *Biochemistry 35*, 691−697.
20. Thompson, P. A., Eaton, W. A., and Hofrichter, J. (1997) *Biochemistry 36*, 9200−9210.
21. Qian, H. (1993) *Biopolymers 33*, 1605−1616.
22. Chakrabartty, A., Kortemme, T., and Baldwin, R. L. (1994) *Protein Sci. 3*, 843−852.
23. Murphy, K. P., Bhakuni, V., Xie, D., and Freire, E. (1992) *J. Mol. Biol. 227*, 293−306.
24. D'Aquino, J. A., Gómez, J., Hilser, V. J., Lee, K. H., Mario Amzel, L., and Freire, E. (1996) *Proteins 25*, 143−156.
25. Jordan, S. R., and Pabo, C. O. (1988) *Science 242*, 893−9.
26. Richmond, T. J., and Richards, F. M. (1978) *J. Mol. Biol. 119*, 537−555.
27. Fersht, A. R. (1985) *Enzyme Structure and Mechanism*; New York, W. H. Freeman and Company.
28. Zwanzig, R. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 148−150.
29. Sosnick, T. R., Mayne, L., Hiller, R., and Englander, S. W. (1994) *Nat. Struct. Biol. 1*, 149−156.
30. Schindler, T., and Schmid, F. X. (1996) *Biochemistry 35*, 16833−16842.
31. Tan, Y.-J., Oliveberg, M., and Fersht, A. R. (1996) *J. Mol. Biol. 264*, 377−389.
32. Lecomte, J., and Matthews, C. R. (1993) *Protein Eng. 6*, 1−10.
33. Fersht, A. R. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 10869−10873.
34. Baldwin, R. L. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 2627−2628.
35. Roder, H. (1997) *Curr. Opin. Struct. Biol. 7*, 15−28.
36. Baldwin, R. L. (1996) *Folding Des. 1*, R1−R8.
37. Sosnick, T. R., Mayne, L., and Englander, S. W. (1996) *Proteins 24*, 413−426.
38. Qian, H., and Schellman, J. A. (1992) *J. Phys. Chem. 96*, 3987−3994.
39. Dyson, H. J., Sayre, J. R., Merutka, G., Shin, H.-C., Lerner, R. A., and Wright, P. E. (1992) *J. Mol. Biol. 226*, 819−835.
40. Minor, D. L., Jr., and Kim, P. S. (1994) *Nature (London) 367*, 660−663.
41. Burton, R. E., Busby, R. S., and Oas, T. G. *J. Biomol. NMR* (In press).

BI980245C