A Minimal Model of Three-State Folding Dynamics of Helical Proteins

Alberto Stizza,[†] Emidio Capriotti,[‡,§] and Mario Compiani*[,§,⊥]

Department of Mathematics and Physics, Catholic University, Brescia, Italy, Department of Physics, University of Bologna, Bologna, Italy, CIRB, Centro Interdipartimentale di Ricerche Biotecnologiche, University of Bologna, Bologna, Italy, and Department of Chemical Sciences, University of Camerino, Camerino, Italy

Received: October 19, 2004; In Final Form: December 22, 2004

A diffusion-collision-like model is proposed for helical proteins with three-state folding dynamics. The model generalizes a previous scheme based on the dynamics of putatively essential parts of the protein (foldons) that was successfully tested on proteins with two-state folding. We show that the extended model, unlike the original one, allows satisfactory calculation of the folding rate and reconstruction of the salient steps of the folding pathway of two proteins with three-state folding (Im7 and p16). The dramatic reduction of variables achieved by focusing on the foldons makes our model a good candidate for a minimal description of the folding process also for three-state folders. Finally, the applicability of the foldon diffusion-collision model to two-state and three-state folders suggests that different folding mechanisms are amenable to conceptually homogeneous descriptions. The implications for a unification of the variety of folding theories so far proposed for helical proteins are discussed in the final discussion.

I. Introduction

Investigations aimed at elucidating the complex features of protein folding have first addressed the calculation or prediction of statical features (essentially, native protein structures[1,2]). This has been depicted as the decipherment of the second half of the genetic code[1] and recently provides the underpinning of the structural genomics project.[3] However, in the past few years increasing evidence has accumulated as to the importance of the dyamical events preceding the stabilization of the native structure. For it has become clear that the process of folding is involved in a delicate regulatory network where perturbations in the timing of parallel events is critical for the correct functioning of the cell or the manifestation of well-known pathologies.[4−6] Moreover, detailed knowledge of the mechanisms underlying the folding of proteins affords guiding principles also in the rational design of proteins.[7] Correspondingly, the focus of research on protein folding has shifted from the study of protein structure to the kinetics and dynamics of the folding process.[2,8−10] However, as of yet, protein folding poses a formidable problem since simulating or monitoring the entire gamut of events that constitute the process is within reach of the available technology only for small peptides.[11]

Much of the inherent complexity of the folding process comes from the frustrated character of proteins.[10] Frustration is the impossibility of optimizing simultaneously the numerous interactions that are dictated on each residue by the native structure of the protein. This feature is responsible for the ruggedness of the underlying energy landscape, which slows down the search for the energy minimum.

A way out of this dead end has emerged since it has been recognized that the undesirable effects of frustration can be partially overcome by a suitable interplay of entropic and energy factors. The key ingredient is the minimal frustration requirement[10] which results in a funnel-shaped landscape that reduces to the right order of magnitude the otherwise astronomical folding times of proteins (Levinthal's paradox).[9,12] Relevant to the present paper is the interpretation of the minimal frustration requirement in helical proteins that was provided in ref 13. This leads to the hypothesis of the existence of regions of the sequence comprising contiguous residues with helical structure, where there is minimal conflict between the constraints imposed by global and local interactions. We refer to these segments as the foldons. This term is not new in the literature[14−16] and its use in this framework was justified in ref 17. In ref 13 we have illustrated a method to detect the minimally frustrated segments with native helical structure directly from sequence.

The comparison with experimental data carried out in ref 13 suggests that the foldons are likely to correspond to the initiation sites of folding where marginally stable elements of secondary structure (IS helices) are formed as the precursors of the native foldon-containing α-helices. This is an important refinement of the assumption, customary in diffusion-collision models,[26] that intrahelical processes are more rapid than the rate-limiting step of folding. We have also argued[13,17] that IS helices are stabilized predominantly by local (or short-range) interactions, i.e., interactions that occur between residues that occupy close positions in sequence. This implies that, upon being formed, they are in a position to induce concomitant reduction of energy and entropy as required by the funnel theory of protein folding.[9,10] Finally, it has been pointed out that the influence of foldons and the IS helices on the folding dynamics extends to longer time scales as they are the building blocks that determine the entire folding dynamics.[17] Evidence for this comes from the foldon diffusion-collision model (henceforth, FDC model), which was applied to small two-state helical proteins in ref 17. The FDC model (see Methods and Theory) provides a quantitative description of folding and is related to the diffusion-collision (DC) model that implements the hierarchical view of protein folding.[18−20]

* Address correspondence to this author. E-mail: mario.compiani@unicam.it.

[†] Catholic University.

[‡] Department of Physics, University of Bologna.

[§] Centro Interdipartimentale di Ricerche Biotecnologiche, University of Bologna.

[⊥] University of Camerino.

10.1021/jp045228d CCC: $30.25
Published on Web 02/15/2005

This paper addresses the paradox arising from the fact that although applicability of hierarchical pictures to three-state folders has been advocated in a recent discussion,[23] and despite the successful application of the FDC model to two-state folders,[17] we have ascertained (see Table 5) that the FDC model grossly fails when applied to proteins Im7 and p16 with three-state folding.[33,42−44] To settle this issue we introduce the FDC3 model as an extension of the FDC model[17] and then show that the FDC3 scheme successully rationalizes the experimental data for Im7 and p16. The plan of the paper is the following. In the Methods and Theory section we illustrate the new features introduced in the FDC3 model to account for the complexity of three-state proteins. In section III we present the results obtained by applying the FDC3 model to three-state proteins Im7 and p16. The refined description of the folding process provided by the FDC3 model is also tested on the two-state protein λ-repressor. In section IV we compare the effectiveness of both dynamical schemes (FDC and FDC3) on two-state and three-state folders and discuss the physical grounds for their complementarity. In the concluding section we discuss the implications of our results for the general theory of folding.

II. Methods and Theory

The general procedures adopted in this paper to study the folding process consist of the FDC model[17] and its generalization devised in the present section (FDC3 method) to better account for the increased complexity of three-state folding dynamics. Before illustrating the details of both models it is convenient to summarize the common architecture of these two computational schemes. To keep the analysis as simple as possible, it suffices to refer to proteins with three IS helices. The essential steps of the FDC and FDC3 folding models are the following: (a) use of a neural network-based predictor to infer the secondary structure from the protein sequence; (b) construction of the information entropy plot starting from the output of the neural network; (c) detection of the foldons from the entropy plot according to the minimal entropy criterion; (d) calculation of the folding probability of the attendant IS helices; (e) calculation of the folding probabilities of aggregates with two IS helices; (f) calculation of the folding time via the diffusion-collision dynamics of the IS helices as described in ref 17; and (g) description of possible paths through the intermediate states allowed by the coalescence mechanism that is inherent in the FDC/FDC3 models. State numeration follows the diagram of Figure 1. The FDC and FDC3 procedures differ as far as step (e) is concerned.

The search for the foldons is based on a feed-forward neural network that is used to predict the secondary structure of helical proteins. We adopt the simplest partition of the space of structures into α and non-α structures. The specifics of the neural network used in this paper are as described in ref 17. The neural network is trained with the error back-propagation algorithm on a database comprising 822 proteins from the PDB. We process the outputs of the neural network to find the position of the foldons and to estimate the probabilities of formation of the corresponding IS helices. This can be done by exploiting the information entropy profile[13] associated with each protein sequence (Figures 3, 6, and 8).

The fundamental principle is the minimal entropy criterion that was stated in ref 13. Basically, two requisites must be met for a prospective foldon: it must possess both a below threshold entropy minimum and an α-helical native structure. The entropy threshold is defined in ref 40.

The FDC model envisages the early formation of marginally stable protostructures (the IS helices) that undergo thermally activated diffusional motions and binary collisions. The subsequent coagulation of the IS helices leads to the progressive formation of clusters (microdomains) of increasing rank. The rank is the number of IS helices composing the microdomain. The birth of a new microdomain at the expense of the older ones with smaller rank hallmarks the transition to a new state along the folding pathway. In a protein with three IS helices the states and the possible pathways are illustrated in Figure 1. Once all the microdomains participate in the globular cluster with the highest possible order, the folding is considered to be complete.

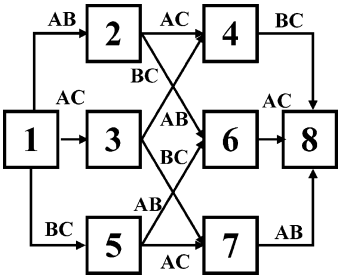

Figure 1. State numbering for the folding process of a protein with three IS helices A, B, and C. State 1 is the unfolded state, where the microdomains of order 1, A, B, and C are completely uncoupled. State 8 represents the native state, where all the native interactions have been established within the aggregate ABC of order 3. The intermediate states correspond to the progressive coalescence of the microdomains. For example, as indicated over the related arrow, passing from state 1 to state 2 requires the aggregation of microdomains A and B (aggregate of order 2). For the further transition to state 4 to occur, microdomain C has to aggregate with A forming an aggregate of order 3. The transitions that result in the establishment of the missing interactions are referred to as internal transitions. By way of example, an internal transition occurs between microdomains B and C on passing from state 4 to state 8. Similar internal transitions are involved on passing from states 6 and 7 to state 8.

The brownian motion of the coupled microdomains is described by the stochastic part of the FDC model that relies on the classical DC theory.[24−28,41] The elementary event is the encounter of two microdomains. The characteristic time τ_{ij} for coalescence of the colliding microdomains (labeled i and j) can be evaluated as

$$\tau_{ij} = \frac{G}{D} + \frac{VL(1 - P_{ij}^{k+l})}{ADP_{ij}^{k+l}} \tag{1}$$

In eq 1, A is the sum of the areas of the colliding microdomains. In the spherical approximation[27,28,41] each pair of microdomains are ascribed the radii R_i and R_j. van der Waals volumes of the helices have been computed by means of TINKER. The radii of the microdomains are evaluated as in ref 27. D is the relative diffusion coefficient defined as $D = k_B T(R_i^{-1} + R_j^{-1})/6\pi\eta$, where η is the viscosity. The temperature T was set to 298 K and the factor $k_B T/6\pi\eta$ was given the value 328.24 Å^2/ns. G and L are defined as

$$G = -\frac{R_{max}^2(5 - 9\epsilon + 5\epsilon^3 - \epsilon^6)}{15\epsilon(1 - \epsilon^3)} \tag{2}$$

$$L^{-1} = \frac{1}{R_{min}} + \alpha\frac{\alpha R_{max}\tanh[\alpha(R_{max} - R_{min})] - 1}{\alpha R_{max} - \tanh[\alpha(R_{max} - R_{min})]} \tag{3}$$

G and L depend on the geometric parameters $R_{min} = R_i + R_j$, $R_{max} = R_{min}$ + linker length and on $\alpha = (D\tau_c)^{-1/2}$, $\epsilon = R_{min}/R_{max}$, and $V = 4\pi(R_{max}^3 - R_{min}^3)/3$. τ_c is the time constant for the coil-helix transition. Collisions are effective when they result

in the aggregation of the colliding *i*th and *j*th microdomains. As far as coalescence is concerned one assumes that effective collision occurs with probability P_{ij}^{k+l}. This implies that with probability1 $- P_{ij}^{k+l}$ no aggregation takes place and the two microdomains separate and start a new diffusional step. If the microdomains *i* and *j* have rank *k* and *l*, respectively, in case an effective collision occurs the resulting more complex microdomain has rank $k + l$. According to the standard DC model[27] we define P_{ij}^{k+l} as the product of the orientational (γ_i^k) and folding (β_i^k) probabilities of the individual microdomains, i.e. $P_{ij}^{k+l} = \gamma_i^k \gamma_j^l \beta_i^k \beta_j^l$. For any microdomain or microdomain cluster γ_i^k is related to the hydrophobic effect, i.e., to the solvation free energy change of the microdomain. Estimates of γ_i^k usually result from computing the ratio of the accessible area lost by the microdomain upon binding to the partner to the total accessible area.[27] The program DSSP provided the accessible surfaces of the various helices as well as the surface that is lost upon contact.

The FDC model departs from the DC model as far as the choice of the microdomains and the evaluation of the folding probabilities β_i^k are concerned. The significant novelty of the FDC model, with respect to the DC model, consists of considering the IS helices (rather than the whole set of native helices) as the critical elementary microdomains that participate in the rate-limiting steps of the folding dynamics. The basic idea of the FDC model is that the folding process can be dissected in semi-independent events taking place in the IS helices that subsequently dictate the kinetics of the whole folding process. The physical rationale for this picture is the quite general fact that proteins, like all complex systems, are characterized by multiscale dynamics.[21] The ensuing hierarchical organization of folding allows us to split the overall dynamics in local fast dynamics and global slow dynamics that are governed respectively by short-range (intra-helical) interactions and long-range (inter-helical) interactions. The fast dynamics pertain to the nucleation and elongation processes of the IS helices that, supposedly, settle very rapidly in a temporary equilibrium state[17] under the dominant influence of short-range forces, consistent with the very definition of a foldon. The regions of sequence that do not belong to the IS helices merely constrain the relative motions of the IS helices. Eventually, one arrives at a simplified description of the protein as a small number of connected beads, each bead being an IS helix.[22] The slow dynamics describe the subsequent formation of the tertiary structure via progressive aggregation of the IS helices. In the later stages of folding, tertiary interactions contribute both to the growth of the coagulated IS helices to their native size and to the formation of the non-IS helices.

The FDC procedure to get estimates of the parameters β_i^1 was introduced in ref 17. The calculation relies on the fact that the entropy profile of the foldons provides information as to the nucleation and elongation constants of the Zimm−Bragg theory applied to the formation of the IS helices. In the FDC model the values of β_i^k are biased as in the original DC model. In particular we have chosen $\beta_i^k = 1$ for transitions with $k > 1$, i.e., transitions involving multihelical aggregates.[27,41] This is a simple representation of the progressive stabilization of multidomains in the later steps of folding due to the increasing momentum of the tertiary interactions.[27,29] The FDC dynamics of the IS helices is successful in computing accurate estimates of the folding times over quite a large range of times.[17] However, on addressing the investigation of more complex folding processes the FDC model outlined above is no longer appropriate, in that for the three-state folders Im7 and p16 we get deviations of nearly an order of magnitude from the experimental folding rates. Experimental and theoretical folding times are compared in Table 5. We assume that the weak point of the FDC model is the approximation $\beta_i^k = 1 \; \forall \; k > 1$, which overemphasizes the self-stabilization properties of higher rank microdomains. Therefore we devise a new procedure for evaluating the β_i^k for aggregates of rank $k \geq 2$. The improved FDC model is henceforth referred to as the FDC3 model. In the cases examined in the present paper it suffices to consider $k = 2$. In an *n*-residue helix let us view $\beta_i^1 = p_1 p_2 ... p_n$ as the product of the probabilities that each individual residue is in the helical (i.e. folded) conformation. Each residue belonging to the *i*th helix can be assigned an average folding probability estimated by means of the following mean field approximation

$$\bar{p} = \sqrt[n]{\beta_i^1} \tag{4}$$

As soon as the *i*th helix interacts with another IS helix, we distinguish among the *R* residues that are definitely stabilized in their helical state from the $n - R$ that keep fluctuating between the helical and the coil state. On computing the new folding probability $\bar{\beta}_i^1 = p_1' p_2' ... p_n'$ we assume that the stabilized residues have $p_k' = 1$ whereas the fluctuating ones have $p_k' = \bar{p}$. To decide a reasonable order of magnitude for *R* we consider that the inter-helical interactions have a stabilizing effect proportional to the relative lost surface (RLS) (ratio of the loss in solvent accessible surface (ΔSAS) to the total solvent accessible surface (SAS)). We assume that the threshold value for the surface loss, L_{thresh} for all the residues to be fixed in the helical state, corresponds to the actual loss as it results from the native tertiary structure of the protein. We have taken $L_{\text{thresh}} = 30\%$ in the case of full alignment of the helices along their axes, and lower values on increasing mutual orientation from parallel to perpendicular (see, for instance, the values chosen for Im7). Thus $L_{\text{thresh}}/n = \text{RLS}/R$, which implies that

$$R = \text{RLS} \frac{n}{L_{\text{thresh}}} \tag{5}$$

where *R* is to be approximated to the nearest integer. In conclusion, the folding probability to be associated with an IS helix within a microdomain of rank 2 is

$$\bar{\beta}_i^1 = \bar{p}^{n-R} \tag{6}$$

where *R* is estimated as in eq 5. Therefore, for the folding probability of the *k*th microdomain of rank 2 composed of IS helices *i* and *j* we assume

$$\beta_k^2 = \bar{\beta}_i^1 \, \bar{\beta}_j^1 \tag{7}$$

where $\bar{\beta}_i^1$ and $\bar{\beta}_j^1$ are computed according to eq 6.

The time evolution of the probabilities of the different states P_i ($i = 1, 2, ..., 8$) is ruled by a master equation in which the transition probabilities per unit time are computed as $(\tau_{ij})^{-1}$.[27] Following refs 17 and 27 we have simplified our simulations by treating the aggregation reactions as if they were irreversible. Accordingly, we have set equal to zero the transition probabilities that describe the dissociation of any microdomain. Calculations of the diffusion-collision processes were performed with MATLAB. The folding time is the time required for the probability of the native state to attain the value 0.6. The folding

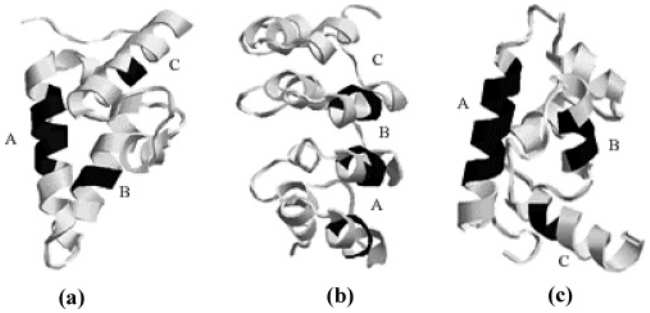

Figure 2. Three-dimensional crystallographic structures and foldons of the two-state and three-state proteins examined in the present paper. Foldons are labeled A, B, and C. (a) Structure of protein Im7. The foldons within the appropriate IS helices are represented as black stretches of native helices 1, 2, and 4. Helix 3 is not predicted by the neural network. (b) Three-dimensional structure of protein p16. The foldons within the appropriate IS helices are represented as black stretches of native helices 2, 3, and 5. The foldon belonging to the native helix 7 is not displayed since it plays a negligible role in the folding dynamics. (c) Three-dimensional structure of the λ-repressor protein. The foldons within the appropriate IS helices are represented as black stretches of native helices 1, 4, and 5. These pictures were obtained with RASTOP 2.0.3.

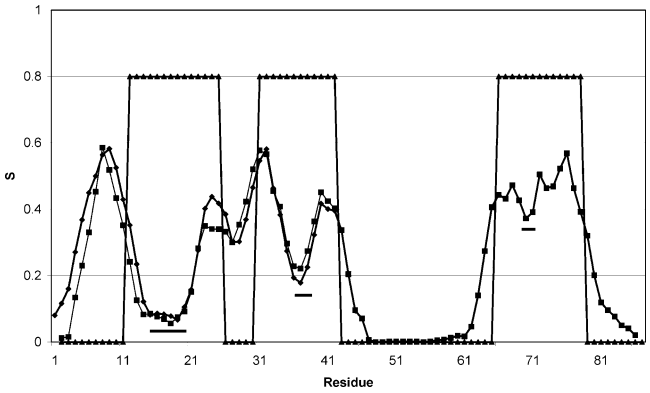

Figure 3. Entropy profiles of proteins Im7 (squares) and Im9 (diamonds) derived according to ref 13. Helical traits predicted by the neural network are marked by the nonzero plateaus of the step function superimposed on the entropy plot. Zero values indicate nonhelical segments. Note that the short native helix 3 (residues 52−55, displayed in Figure 2a) is not predicted by the neural network. Black bars mark foldons A (13−19), B (35−36), and C (69−70) (see Table 1) belonging respectively to native helices 1, 2, and 4. Despite moderate sequence homology, the two curves are remarkably similar. For the sake of readibility of the diagram predicted helical traits of Im9 are not shown. As in the case of Im7 the native helix 3 (residues 52−55) is not predicted by the neural network. Foldons A (15−20), B (36−38), and C (70−71) of Im9, belonging respectively to native helices 1, 2, and 4, are not marked since they considerably overlap with the foldons of Im7.

times of Table 5 correspond, for example, to the time τ_{FDC3} such that $P_8(\tau_{FDC3}) = 0.6$ in the graphs of Figures 4, 5, 7, and 9.

III. Results

A. The Folding of Im7. Im7 (PDB code, 1CEI) is an helical immunity protein with four native helices[33] (Figure 2a). The secondary structure of 85 out of the 87 residues of 1CEI have been resolved crystallographically. The entropy profile (Figure 3) was calculated following the procedure of ref 13 (see Methods and Theory). According to the minimal entropy criterion[13,40] only helices 1, 2, and 4 meet the requirements for being considered IS helices. In the sequel we label them microdomains A, B, and C, whereas ordinals are used to denote the native helices. The location and size of foldons as predicted by the neural network are shown in Figures 2a and 3. The relevant parameters for the implementation of the FDC3 calculation are gathered in Table 1.

Helix 3 (residues 52−55 of the PDB file) is not predicted by the neural network (Figure 3). Accordingly, it is not subjected to the minimal entropy criterion for the search of foldons and the IS helices. This is a typical failure of the generalization capability of the neural network that, due to the low entropy value assigned by the neural network-based predictor (Figure 3), can be ascribed to the paucity of examples of the same type in the database rather than to the noise affecting the sequence-secondary structure mapping.[32]

As we shall see, misprediction of helix 3 has little effect on the calculation of the folding time. Visual inspection of the mutual orientations of the IS helices from the three-dimensional structure of Figure 2a shows that helix B is approximately parallel to helix C and angle(A,C) > angle(A,B). This suggests $L_{thresh}(B,C) = 30\%$, $L_{thresh}(A,C) = 15\%$, and $L_{thresh}(A,B) = 20\%$ (see Methods and Theory) to be used in eq 5. The geometrical parameters needed to evaluate the folding time (eq 1) are displayed in Table 1. For the characteristic time τ_c in eq 3 we chose the value $\tau_c = 1$ ns, that lies in the usual range of values used in the DC and FDC models.[17,27,28]

Valuable insight into the actual sequence of states traversed by the protein during the folding process comes from the curves of Figure 4, that displays the probabilities in time of the relevant states, according to the FDC3 computation and, for comparison, to the original FDC scheme.

As shown also in Table 5, the folding time estimated by the FDC model is 0.77 ms, which is sensibly smaller than the experimental time 3.06 ms,[34] whereas the result obtained with the FDC3 model, $\tau_{FDC3} = 2.98$ ms, compares much better with the experimental value. Notably, different values of the folding time were reported in the literature (4.20 ms, using the kinetic data of ref 33 and 3.36 ms with 0.4 M Na_2SO_4[34]). It is evident that the more realistic description of the interactions of microdomains of rank 2 has a visible effect on the overall kinetics of the folding process of Im7.

In parallel, the lifetimes of states 2 and 4 increase on passing from the FDC to the FDC3 scenario. More precisely, states 2 and 4 show slower decay in the FDC3 calculations and this reflects the experimental observability of the intermediate state.[33] States 2 and 4 seem to reach a temporary pseudoequilibrium whereas their decline accompanies the rise of state 8. State 8 corresponds to the native state where the missing coupling of microdomains B and C is eventually established. The putative role of states 2 and 4 in the intermediate of folding indicates that helices 1, 2, and 4 are mainly involved in the critical step of the process. Our results are consistent with experimental data[33] indicating that the on-pathway intermediate of Im7 includes only the three IS helices detected by the neural network (Figures 2a and 3). Also the mutation study carried out in ref 33 indicates that helix 3 is likely to participate in the helical core of Im7 only after the aggregate composed of helices A, B, and C has formed. This is in accord with the fact that helix 3 is not expected to belong in the set of the IS helices.

Next, in experimental works rapid preequilibrium is assumed between the unfolded state and the intermediate state, followed by slow relaxation to the native state. The curves of Figure 4 do show the same features.

B. The Folding of Im9. The kinetics of the 86 residues protein Im 9 (PDB code, 1IMQ) also has been experimentally characterized.[34] It is commonly accepted that the folding mechanism is conserved within the same family of proteins,[36−39] but Im7 and Im9 seem to be an exception. As a matter of fact,

TABLE 1: Geometrical and Stability Parameters of Im7[a]

	A	B	C		A	B	C
l (Å)	$l_{AB} = 17.5$	$l_{BC} = 80.5$		β^1	0.062	0.036	0.019
Δn	$\Delta n_{AB} = 5$	$\Delta n_{BC} = 23$		n	15	12	13
r (Å)	8.58	7.85	8.15	foldons	13−19	35−36	69−70
V (Å^3)	1958	1501	1678	$\bar{p}$	0.831	0.758	0.737
SAS (Å^2)	1745	1410	1681				

	A			B			C		
	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$
A				163 (9%)	7	0.227	235 (13%)	13	0.690
B	184 (14%)	8	0.330				15 (1%)	1	0.048
C	243 (15%)	12	0.737	15 (1%)	1	0.026			

[a] Shown are the geometrical and stability parameters of microdomains A, B, and C (upper part) and features of the microdomains of rank $k = 2$ of Im7 (in matrix form in the lower part of the table). l represents the length of the linker joining microdomains i and j (i, j = A, B, C). Δn is the number of residues composing the linker of length l. Note that $l = 3.5\Delta n$, since each residue has the average length of 3.5 Å. r denotes the radius of the individual microdomains. V is the volume of the microdomain at hand. SAS is the solvent accessible surface. The folding probabilities β_i^1 are estimated graphically from the entropy profile of Figure 3. n is the number of residues in the native helices. The scopes of the three foldons are given specifying the numbers of the first and last residues of each of them. $\bar{p}$ is the average probability per residue calculated as in eq 4. The lower part of the table collects in matrix form the lost surface ΔSAS, the relative lost surface (RLS) and the new folding probability $\bar{\beta}_k^1$. The number of native residues in each IS helix, R, is computed as in eq 5. The probabilities $\bar{\beta}_j^1$ are calculated according to eq 6. The entries refer to helix Y (Y = A, B, C, matrix row) after binding with helix X (X = A, B, C, matrix column). These values are then used to calculate the folding probability β_k^2 as in eq 7. IS helices B and C lose a small surface upon contact which we fix to 1%.

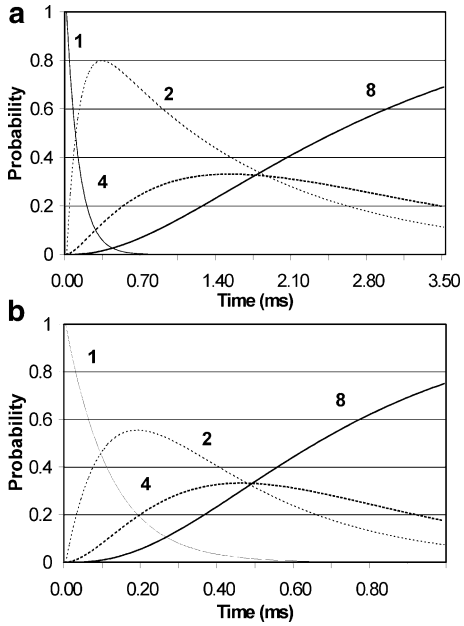

Figure 4. Behavior in time of the probabilities of the critical states of protein Im7. Times are expressed in milliseconds. States with populations less than 0.01 are not shown. State 1 is the denatured state and state 8 is the native state (Figure 1). In the intermediate states 2 and 4 only partial aggregation of the microdomains has occurred. In state 2 microdomains A and B have coalesced to form the precursor of the observable intermediate state 4 in which A interacts simultaneously with B and C. The curves of panel (a) are calculated by means of the FDC3 model. For comparison we report in panel (b) also the curves calculated by means of the FDC model. State 2 exhibits a detectable population both in the FDC and in the FDC3 model but has a faster decay in the FDC plot. Also state 4 becomes more visible in the FDC3 calculation due to its longer lifetime. Note the different time ranges on the abscissas of panels (a) and (b). All the other possible intermediate states play a negligible role. State 1 exhibits practically the same features in the FDC and in the FDC3 calculations.

Im9 is a two-state folder while Im7 has a three-state folding dynamics. Therefore Im9 is a critical benchmark as it is interesting to check whether the FDC and FDC3 models succeed in perceiving such a difference in the folding process. In Figure 3 we have superimposed the entropy profiles of Im 7 and Im 9. Despite moderate sequence homology (60%) the entropy plots of Im7 and Im9 are strikingly similar. The foldons and the helices of Im9 are almost coincident with those of Im7 (Figure 3 and Table 2) apart from modest differences within the regions of foldons A and B that are the major ones responsible for the modified kinetics of folding. The relevant parameters for computing the folding probabilities of the microdomains and eventually to evaluate the folding time (eq 1) are reported in Table 2. In the FDC3 calculation we used the same time τ_c and the same thresholds L_{thresh} as for Im7, due to the similarity of the tertiary structure of Im9 (not shown) to that of Im7 (Figure 2a).

The folding pathway of Im9 resembles closely that of Im7. Similarly, states 2 and 4 are traversed with highest probability (Figure 5). However, meaningful differences emerge as to the behavior in time of the populations of these two states predicted by the FDC and the FDC3 schemes. State 2 of Im9 preserves approximately the same kinetics already calculated for Im7 save for moderate changes in the maximal population. By contrast, state 4 of Im9 is much less populated than state 4 of Im7 both in the FDC and FDC3 calculations. Actually, state 2 of Im9 is the only intermediate state to be significantly populated. As far as the different trend predicted by the FDC and the FDC3 scenarios is concerned, state 2 grows approximately at the same rate and attains more or less the same population according to both calculations but decays more slowly according to the FDC3 method. On passing from the FDC to the FDC3 procedure state 4 of Im9 undergoes minor changes in amplitude and decay rate. We claim that this finding is consistent with the absence of experimentally detectable intermediates in the folding process of Im9, albeit the intermediate 2 is visible in Figure 5. The relationships between FDC/FDC3 aggregates (Figure 1) and the intermediates studied by the experimentalists are clarified in the Discussion.

C. The Folding of Protein p16. Protein $p16^{INK4a}$ (p16 for brevity, PDB code 1BI7) is a member of the INK4 family.[42] Kinetic data have been reported recently.[42−44] p16 has 154 residues and seven native helices[43] (Figure 2b). The PDB file covers the region from residue 10 to residue 134. The entropy profile is drawn in Figure 6. The residue numbering used in the present paper, being based on the PDB file, is shifted by 9 amino acids with respect to the numbering of the entire sequence

TABLE 2: Geometrical and Stability Parameters of Im9[a]

	A	B	C		A	B	C
l (Å)	$l_{AB} = 17.5$	$l_{BC} = 80.5$		β^1	0.087	0.019	0.008
Δn	$\Delta n_{AB} = 5$	$\Delta n_{BC} = 23$		n	14	12	13
r (Å)	7.34	7.20	7.16	foldons	15−20	36−38	70−71
V (Å^3)	1664	1575	1549	$\bar{p}$	0.850	0.719	0.690
SAS (Å^2)	1544	1725	1505				

	A			B			C		
	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$
A				289 (19%)	9	0.444	224 (15%)	14	1.000
B	267 (15%)	6	0.138				76 (4%)	2	0.037
C	245 (16%)	13	1.000	59 (4%)	2	0.017			

[a] Shown are the gometrical and stability parameters of microdomains A, B, and C (upper part) and features of the microdomains of rank $k = 2$ of Im9 (in matrix form in the lower part of the table). l represents the length of the linker joining microdomains i and j (i, j = A, B, C). Δn is the number of residues composing the linker of length l. Note that $l = 3.5\Delta n$, since each residue has the average length of 3.5 Å. r denotes the radius of the individual microdomains. V is the volume of the microdomain at hand. SAS is the solvent accessible surface. The folding probabilities β_i^1 are estimated graphically from the entropy profile of Figure 3. n is the number of residues in the native helices. The scopes of the three foldons are given specifying the numbers of the first and last residues of each of them. $\bar{p}$ is the average probability per residue calculated as in eq 4. The lower part of the table collects in matrix form the lost surface ΔSAS, the relative lost surface (RLS), and the new folding probability $\bar{\beta}_k^1$. The number of native residues in each IS helix, R, is computed as in eq 5. The probabilities $\bar{\beta}_j^1$ are calculated according to eq 6. The entries refer to helix Y (Y = A, B, C, matrix row) after binding with helix X (X = A, B, C, matrix column). These values are then used to calculate the folding probability β_k^2 as in eq 7.

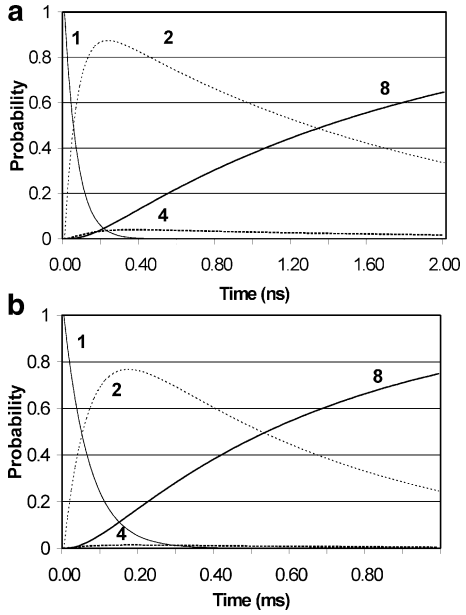

Figure 5. Behavior in time of the probabilities of the critical states of protein Im9. Times are expressed in milliseconds. States with populations less than 0.01 are not shown. State 1 is the denatured state and state 8 is the native state (Figure 1). The curves of panel (a) are calculated by means of the FDC3 model. For comparison we report in panel (b) also the curves calculated by means of the FDC model. As in Figure 4, the intermediate states 2 and 4 play a fundamental role. In state 2 microdomains A and B coalesce to form the precursor of the observable intermediate state 4 in which A interacts simultaneously with B and C. State 4 of Im9 exhibits a detectable population both in the FDC and in the FDC3 model but has a shorter lifetime in the FDC than in the FDC3 diagram (note the different time ranges on the abscissas of panels (a) and (b)). State 4 is seen as being scarcely populated both in the FDC and in the FDC3 model. State 2 preserves approximately the same features in Im7 and Im9. All the other possible intermediate states play a negligible role. State 1 exhibits practically the same features in the FDC and in the FDC3 calculations.

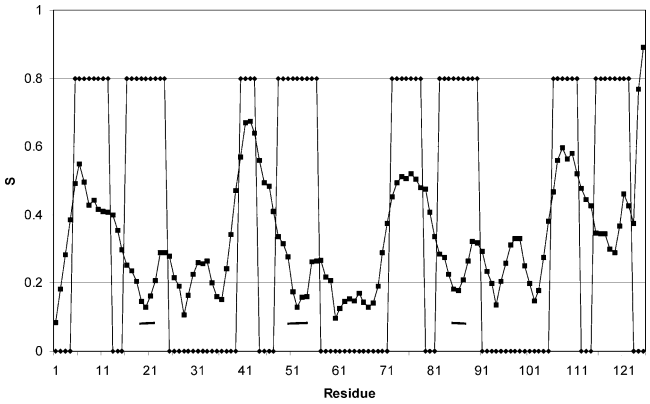

Figure 6. Entropy profile of protein p16 obtained following ref 13. Predicted helical traits are marked by the nonzero plateaus of the step functions superimposed on the entropy plot. Zero values indicate nonhelical segments. Helices 2, 3, 5, and 7, out of seven native helices, contain a foldon. Helix 7 fulfills the requirements for being considered as an IS helix but is not considered in the overall dynamics (see text) so that the three IS helices relevant to the FDC/FDC3 calculations belong to the native helices 2, 3, and 5. Black bars mark the foldons within the three IS helices A, B, and C. The foldons associated with IS helices A, B, and C span the following segments: foldon A (19−21), foldon B (51−54), and foldon C (85−87). The modular structure of p16, made up of four ANK repeats,[43] is visible from the approximate symmetries of the entropy profile. The plot has an approximate periodicity of 65 residues, since that region is the junction of the two halves of the protein (connecting the second with the third repeat).

(used for example in ref 44). Helices 2, 3, 5, and 7 are eligible as IS helices. We surmise that helix 7 has a negligible role in the folding dynamics due to the very small contact area with the other three foldons. Therefore, consistent with the mechanism of diffusion-collision, we use only the IS helices hosted within the native helices 2, 3, and 5 as the fundamental determinants of the dynamics (a similar approximation was made in the analysis of the villin headpiece subdomain[30]).

Interestingly, the relevance of our foldons in the rate-limiting step of the process is confirmed by the estimates of the ϕ-values in the transition state reported in ref 44. More specifically, ϕ-values close to 1 are assigned to A21 and V86 (our numbering) that fall at the end of foldons A and C (Table 3). The further ϕ-value close to 1 assigned to A59 gives less precise evidence since it falls a bit outside of foldon B. However, this discrepancy may well be due to the noise that affects the entropy plot and in particular the determination of the ends of the predicted helices.[32] Finally, residues V117 and A118 do belong to the last foldon predicted by the neural network in the last IS

TABLE 3: Geometrical and Stability Parameters of p16[a]

	A	B	C		A	B	C
l (Å)	$l_{AB} = 80.5$	$l_{BC} = 87.5$		β^1	0.020	0.045	0.024
Δn	$\Delta n_{AB} = 23$	$\Delta n_{BC} = 25$		n	9	9	9
r (Å)	7.21	7.14	7.22	foldons	19−21	51−54	85−87
V (Å^3)	1163	1130	1166	$\bar{p}$	0.647	0.708	0.661
SAS (Å^2)	1142	1158	1171				

	A			B			C		
	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$
A				206 (18%)	5	0.176	6 (0.5%)	1	0.031
B	247 (21%)	6	0.356				185 (16%)	5	0.252
C	6 (0.5%)	1	0.036	193 (16%)	5	0.191			

[a] Shown are the geometrical and stability parameters of microdomains A, B, and C (upper part) and features of the microdomains of rank $k = 2$ of protein p16 (in matrix form in the lower part of the table). l represents the length of the linker joining microdomains i and j (i, j = A, B, C). Δn is the number of residues composing the linker of length l. Note that $l = 3.5\Delta n$, since each residue has the average length of 3.5 Å. r denotes the radius of the individual microdomains. V is the volume of the microdomain at hand. SAS is the solvent accessible surface. The folding probabilities β_i^1 are estimated graphically from the entropy profile of Figure 6. n is the number of residues in the native helices. The scopes of the three foldons are given specifying the numbers of the first and last residues of each of them. $\bar{p}$ is the average probability per residue calculated as in eq 4. The lower part of the table collects in matrix form the lost surface ΔSAS, the relative lost surface (RLS), and the new folding probability $\bar{\beta}_k^1$. The number of native residues in each IS helix, R, is computed as in eq 5. The probabilities $\bar{\beta}_j^1$ are calculated according to eq 6. The entries refer to helix Y (Y = A, B, C, matrix row) after binding with helix X (X = A, B, C, matrix column). These values are then used to calculate the folding probability β_k^2 as in eq 7.

TABLE 4: Geometrical and Stability Parameters of the λ-Repressor[a]

	A	B	C		A	B	C
l (Å)	$l_{AB} = 91.0$	$l_{BC} = 28.0$		β^1	0.099	0.085	0.048
Δn	$\Delta n_{AB} = 26$	$\Delta n_{BC} = 8$		n	18	12	12
r (Å)	9.26	7.75	7.93	foldons	13−22	63−66	82−83
V (Å^3)	2467	1445	1578	$\bar{p}$	0.879	0.814	0.776
SAS (Å^2)	2307	1378	1475				

	A			B			C		
	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$	ΔSAS (RLS)	R	$\bar{\beta}_j^1$
A				151 (7%)	4	0.164	67 (3%)	10	0.405
B	157 (11%)	5	0.237				115 (8%)	12	1.000
C	76 (5%)	12	1.000	99 (7%)	12	1.000			

[a] Shown are the geometrical and stability parameters of microdomains A, B, and C of the λ-repressor (upper part) and features of the microdomains of rank $k = 2$ of the same protein (in matrix form in the lower part of the table). l represents the length of the linker joining microdomains i and j (i, j = A, B, C). Δn is the number of residues composing the linker of length l. Note that $l = 3.5\Delta n$, since each residue has the average length of 3.5 Å. $\Delta n_{AB} = 26$ results from shortening the linker in order to account for the boundary between helix 3 and 4 that is mispredicted by the neural network. Actually, the neural network predicts a unique helical segment spanning crystallographic helices 3 and 4 (see Figure 8). However, as visible in Figure 8, residues 46−51 belong to an α-helix whose entropy minimum is very close to the threshold for being classified as an IS helix. In this case, to counterbalance the effect of noise affecting the entropy signal (see Discussion), we use the crystallographic boundaries for helices 3 and 4 (shown in Figure 8) and have renormalized the actual loop length $\Delta n_{AB} = 32$ to 26 to account for the loop shortening due to the possible formation of helix 3. This correction is much in the same spirit of the determination of the effective protein length suggested in ref 50. It optimizes the predition of the folding rate of the λ-repressor (see Discussion). r denotes the radius of the individual microdomains. V is the volume of the microdomain at hand. SAS is the solvent accessible surface. The folding probabilities β^1 are estimated graphically from the entropy profile of Figure 8. n is the number of residues in the native helices. The scopes of the three foldons are given specifying the numbers of the first and last residues of each of them. $\bar{p}$ is the average probability per residue calculated as in eq 4. The lower part of the table collects in matrix form the lost surface ΔSAS, the relative lost surface (RLS), and the new folding probability $\bar{\beta}_k^1$. The number of native residues in each IS helix, R, is computed as in eq 5. The probabilities $\bar{\beta}_j^1$ are calculated according to eq 6. The entries refer to helix Y (Y = A, B, C, matrix row) after binding with helix X (X = A, B, C, matrix column). These values are then used to calculate the folding probability β_k^2 as in eq 7.

helix (residues 114−120). Using the same notation of the previous cases, we label the relevant foldons and the corresponding IS helices as A, B, and C. The foldons include the following stretches of the protein sequence: foldon A (19−21), foldon B (51−54), and foldon C (85−87) (see Figure 6). The geometrical parameters needed to evaluate the folding time (eq 1) are displayed in Table 3. For the characteristic time τ_c in eq 3 we used the value $\tau_c = 1$ ns.

The three-dimensional structure of Figure 2b shows that the native helices are approximately parallel. This suggests the value $L_{thresh} = 30\%$ (see Methods and Theory) to be used in eq 5.

The experimental folding time of p16 is 30.3 ms.[42] The FDC3 time is 27.9 ms whereas the FDC time amounts to 2.67 ms (see Table 5). The FDC3 model is successful in reproducing the right order of magnitude of the folding time and accounts also for other details of the process that are missed otherwise by the FDC model. The striking feature of Figure 7 is the drastic change of the population and decay rate of state 6 as we switch from the FDC to the FDC3 scenario. Panel (a) of Figure 7 shows that state 6 is well populated right after the decay of states 1, 2, and 5. Moreover its population is depleted at a rate that is smaller than predicted by the FDC model (Figure 7b). States 2 and 5 play a minor role in the overall folding process. They are much less persistent and populated than state 6 and essentially serve to channel population into state 6 (see Figure 1). State 6 is a good candidate for being involved in the intermediate of folding I responsible for the three-state dynamics of p16. Recent experimental characterization of state I[42] suggests a rapid

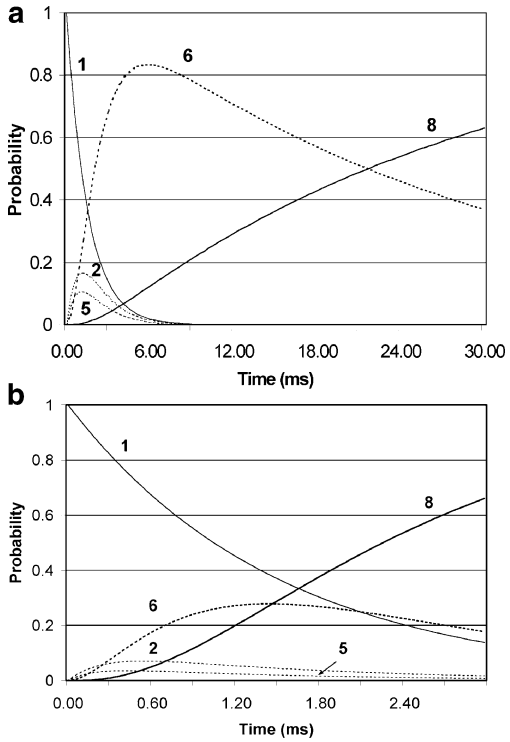

Figure 7. Behavior in time of the probabilities of the critical states of protein p16. Times are expressed in milliseconds. States with populations less than 0.01 are not shown. State 1 is the denatured state and state 8 is the native state (Figure 1). In the intermediate states 2, 5, and 6 only partial aggregation of the microdomains has occurred. In state 2 microdomains A and B coalesce to form the precursor of the observable intermediate state 6. In state 6 microdomain B, besides interacting with A, begins to interact also with C. Alternatively, state 6 is preceded by state 5 in which helix B first binds to helix C. (a) Probabilities of the relevant states computed through the FDC3 model. (b) For comparison we report also the curves calculated by means of the FDC model. The more realistic description of the interactions of microdomains of rank 2 in panel (a) has a visible effect on the overall kinetics of the folding process. The decay of state 1 takes place on the same time scale as the rise and decay of states 2 and 5 both in the FDC and in the FDC3 model. The typical time is about 5 ms (note the different units on the abscissas of panels (a) and (b)). The major changes are visible in the curves of states 6 and 8. In particular, the population of state 6 increases and the FDC3 estimate for the lifetime is approximately 10-fold longer than the FDC value. Correspondingly, state 8 is populated at a rate that is 1 order of magnitude larger in the FDC3 model than in the FDC scenario. Also the folding time increases by 1 order of magnitude (see the text and Table 5). The FDC calculation predicts smaller populations for states 2, 5, and 6.

formation of the intermediate that is not detectable, due to the relatively long dead-time of the stopped-flow apparatus. Experiments also show that the rate-limiting step is the transition I → N, leading to the native state 8. All these features are consistent with the fast rise of the population of state 6 and the concomitant fast decay of states 1, and 5 in the FDC3 plot of Figure 7a.

Comparing the plots computed according to the FDC or the FDC3 model (Figure 7) it is apparent that states 2, 5, and 6 are more visible in the FDC3 calculation. Among them, state 6 undergoes the largest amplification in amplitude and its persistence is increased by approximately 1 order of magnitude. State 1 seems to undergo no significant changes in the FDC3 method as compared to the FDC calculation.

D. The Folding of the λ-Repressor. The λ-repressor protein (PDB code, 1LMB4) is a two-state folder whose folding kinetics has been successfully modeled within the standard DC model[29] and the FDC model.[17] This protein provides also a cogent benchmark for the effectivity of the FDC model in predicting

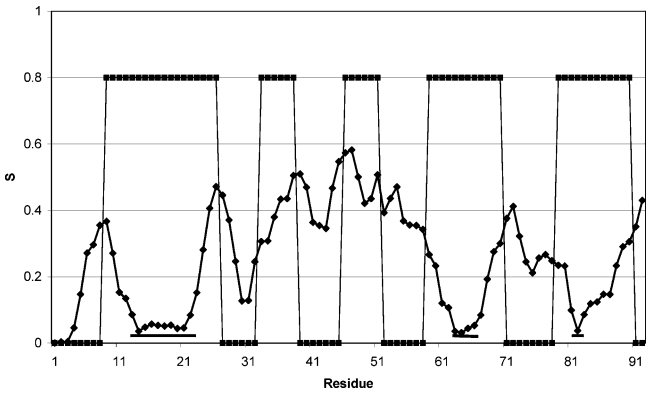

Figure 8. Entropy profile of the λ-repressor derived following ref 13. Predicted helical traits are marked by the nonzero plateaus of the step function superimposed on the entropy plot. Zero values indicate nonhelical segments. Helices 2 and 3 are non-IS helices. The foldons associated with helices 1, 4, and 5 span the following segments (black bars): foldon A (13−22), foldon B (63−66), and foldon C (82−83). The boundary between the native helices 3 and 4 is missed by the neural network, which predicts a unique helical region encompassing helices 3 and 4. The step function in correspondence of helices 3 and 4 represents the crystallographic data. The entropy plot indicates that helix 3 is a quasi-IS helix.

TABLE 5: Comparison of Computed and Experimental Folding Times[a]

protein	folding	τ_{FDC} (ms)	τ_{FDC3} (ms)	τ_{exp} (ms)
Im7	three-state	0.77	2.98	3.06
p16	three-state	2.67	27.90	30.30
λ-repressor	two-state	0.21	0.27	0.20
Im9	two-state	0.68	1.78	0.67

[a] As far as the computed times are concerned, τ_{FDC} values have been calculated according to the FDC model,[17] whereas the τ_{FDC3} values are evaluated following the FDC3 model (see Methods and Theory). τ_{FDC} of Im9 and the λ-repressor have been computed first in ref 17. Note that the folding time of Im7 is 3.06 with 4 M Na_2SO_4.[34] The experimental times τ_{exp} for Im7 and Im9 have been computed by using the kinetic data of refs 33 and eq 6 of ref 34. For the λ-repressor the data have been drawn from ref 45.

the change of the folding rate on point mutations of the wild type (M. Compiani et al., in preparation). Like Im9, the λ-repressor is used here as a term of comparison to test the performance of the FDC3 model in a case where it is not strictly needed. The λ-repressor is a five-helix bundle whose N-terminal fragment 1−95 is considered in the present investigation (Figure 2c). From its entropy plot (Figure 8) it is apparent that only three helical segments are eligible as IS helices (helices 1, 4, and 5) which we label A, B, and C. Inspection of the tertiary structure shows that IS helices B and C as well as A and C are approximately mutually perpendicular. As argued in the Methods and Theory, in these two cases the appropriate threshold to be used in eq 5 is $L_{thresh} = 5\%$. The threshold $L_{thresh} = 30\%$ was assumed for the AB microdomain in which the coalescing IS helices are approximately parallel. The relevant parameters for computing the folding probabilities of the microdomains and eventually to evaluate the folding time (eq 1) are reported in Table 4. For the characteristic time τ_c in eq 3 we used the value $\tau_c = 0.1$ ns.[17]

Useful insights into the details of the kinetics of folding are readily obtained from the FDC and FDC3 models. The relevant diagrams for the λ-repressor are plotted in Figure 9.

The experimental folding time is 0.20 ms[29] whereas the FDC3 time is 0.27 ms. The FDC time amounts to 0.21 ms (see Table 5).

From the FDC calculation only state 5 emerges with a moderate population. The FDC3 scheme shows two alternative

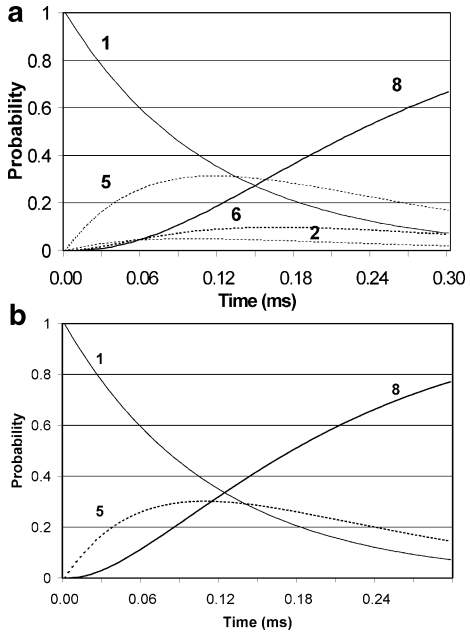

Figure 9. Behavior in time of the probabilities of the critical states of the two-state λ-repressor protein. Times are expressed in milliseconds. States with population less than 0.01 are not shown. State 1 (denatured state) and state 8 (native state) are the only states that are populated. (a) Probabilities of the relevant states computed through the FDC model. (b) Probabilities of the relevant states computed through the FDC3 model. The more detailed description of the interactions of the microdomains of rank 2 in panel (b) has a moderate effect on the overall kinetics of the folding process (see Table 5). Intermediate state 5 is assigned comparable populations and decay both in the FDC and in the FDC3 calculations.

paths 1−2−6−8 and 1−5−6−8 (Figure 1). The first one is by far dominant since state 5 has a much larger population than state 2. State 6, where the two paths merge, is also scarcely populated in the FDC3 scheme and is not at all visible in the FDC calculation (i.e. below the 0.01 threshold).

IV. Discussion

A general overview of the results obtained with the FDC and the FDC3 models for the folding times of the two-state and three-state folders examined in this paper is given in Table 5. The fundamental outcome is that the FDC3 model succeeds in reproducing the experimental folding rates of Im7 and p16 that are instead poorly accounted for by the FDC scheme. However, the successful test of the FDC3 model has additional implications that can be better understood by considering the physical meaning of the FDC3 procedure (outlined in the Methods and Theory section).

The physically grounded prescription for the evaluation of the β_i^k coefficients with $k \geq 2$ (see Methods and Theory) is the major novelty introduced through the FDC3 calculation, since so far these variables were set to sensible but otherwise arbitrary values within the framework of the DC or FDC model.[17,26−28,30] The proteins examined in this paper do possess three foldons, so that it suffices here to apply the new calculation to the β_i^2 (folding probability of the *i*th microdomain of order 2) to supplant the FDC and DC prescription $\beta_i^2 = 1$. The FDC3 estimates $\beta_i^2 \leq 1$ are computed according to eqs 6 and 7.

The FDC3 scheme introduces an additional coupling between formation of partial and local structures and more global structures, besides the similar coupling which is already introduced by the orientational γ coefficients in the FDC and DC methods (Methods and Theory). Indeed, also the FDC3 estimates for β_i^2 depend on the three-dimensional packing through the solvent accessible surface area (RLS in eq 5) lost upon aggregation and subsequent formation of a microdomain of higher order. This indicates that in the FDC3 model local and global events get more intertwined than in the FDC or DC models. As a matter of fact the DC assumption $\beta_i^2 = 1$ means that all the native helices are stabilized once they get bound to any aggregate of order $k \geq 2$. This amounts to saying that a minimal development of the tertiary structure is sufficient to ensure the proper folding of the native helices involved in the current collision.

Instead, within the FDC model, the IS helices and the non-IS helices have different fates. Actually, formation of the non-IS helices does not take place in the early stages of the folding process but requires necessarily that a more developed approximation of the native scaffold is formed, with the dominant contribution of the IS helices. In the FDC3 model this feature is somewhat more evident since also the aggregates of the IS helices of order $k = 2$ are possibly assigned marginal native character since $\beta_i^2 < 1$ (eqs 6 and 7). This implies that the IS helices themselves need more assistance from the tertiary structure being formed to enter the native folded state.

This aspect stresses the coexistence of cooperative and hierarchical features in the DC, FDC, and FDC3 models that seem capable of accounting for different degrees of cooperativity and modularity of the folding mechanism. If we accept the view[18,35] that the fully hierarchical and the fully cooperative mechanisms are limiting scenarios of a continuous range of different mechanisms, we can argue that these limiting cases can be approached by the FDC model by tuning properly its variables. In accord with the remarks of the previous paragraph, we can conclude that by turning from the FDC to the DC calculation (by performing the limit $\beta_i^2 \rightarrow 1$ and dropping any discrimination between IS and non-IS helices) the FDC model becomes more hierarchical, as formation of the native secondary structure is more independent of the stabilization of the native tertiary structure. On the contrary, by introducing the FDC3 estimates ($\beta_i^2 \leq 1$), the FDC folding mechanism becomes more cooperative since stabilization of local and global structures tends to become more dependent from each other. This item is discussed in more quantitative terms in by M. Compiani (submitted for publication). In light of these considerations the successful application of the FDC3 model to Im7 and p16 implies that hierarchical features inherent in the FDC3 model are also found in folding mechanisms having higher complexity than the bare two-state mechanisms described by the FDC model in ref 17. This is in agreement with the conclusions of ref 23 that discusses the possibility of extending the modular view of folding to three-states folders.

Two-state folders have been included in our analysis to provide a more complete test of the FDC and FDC3 models on folding processes of different complexity. Expectedly, the folding times for the proteins examined in this paper obey the relation $\tau_{FDC} < \tau_{FDC3}$. This is clearly ascribed to the overestimation of the folding probabilities $\beta_j^2 = 1$ that is proper to the FDC model. From Table 5 it is apparent that the improved calculation of the folding probabilities for complex microdomains of rank $k = 2$ is crucial for the three-state proteins, but is by far less critical for the two-state proteins λ-repressor and Im9. For the latter proteins, the FDC3 values are expected to converge to the FDC folding rates. The data in Table 5 conform satisfactorily to this prediction.

The study of the homologous proteins Im7 and Im9 is also quite illuminating. The degeneracy of the folding code is well

exemplified in the case of Im7 and Im9 in which different sequences give rise to similar structures. In addition, albeit sequence homology of Im7 and Im9 is moderate, the two proteins have quite similar entropy plots (Figure 3), although Im7 has a three-state mechanism and Im9 has a two-state folding process. For this reason Im7 and Im9 are quite a critical benchmark for the FDC and FDC3 models. The fact that the folding rates of both proteins are well reproduced confirms the pivotal role played by the foldons, such that appropriate modulation of the entropy profile in the foldon regions can make the folding mechanism switch from two-state to three-state mode.

In general, the effectivity of the FDC and FDC3 models in predicting the folding times of proteins (see Table 5 and the additional data in ref 17) confirms that the foldons are the key elements of folding. Such a possibility of focusing on a limited number of IS helices as determinants of the folding dynamics supports the contention that the FDC scenario is a promising tool for building minimal models of protein folding that are nonetheless sensitive to sequence features. In this respect the neural network is responsible for the sequence specific character of both models that makes them effective in capturing the properties of relatively short stretches of the residue sequence. A case in point is the discrimination of IS and non-IS helices within an individual protein. The peculiar sensitivity of the FDC/FDC3 models to as small details of sequence as single point mutations will be examined by M. Compiani (in preparation).

In the framework of the FDC and FDC3 models the IS helices act as building blocks that resemble the elementary units used in other approaches to carry out a dissection of folding on a thermodynamic or structural basis.[22,31,46] In ref 17 we have shown how to derive the thermodynamic properties of the IS helices from the sequence of the protein. Interestingly, our estimates of the stability of the IS helices correlate well with the experimental helicities measured in the same isolated peptides.[17] Focusing on the IS helices affords a gross-grained partition of the phase space that is based on all the possible pairings of the available units (see Figure 1). An interesting feature of the FDC and FDC3 models is that through the neural network one takes maximal advantage of the pieces of information coded in the residue sequence. This is progress toward theoretical mehods that are capable of predicting folding rates from sequence. As already stressed above, unlike recent methods that rely entirely on sequence information,[49,50] the FDC model uses also features derived from the tertiary structure (see Methods and Theory). However, the FDC method gives a deeper insight into the details of protein folding (e.g. the nature of possible intermediates) that is useful to devise modified computational schemes such as the FDC3 method proposed in this paper.

In this connection the possibility to visualize and estimate the lifetimes and populations of the intermediates as well as the relative probability of different pathways is a notable feature of the DC/FDC/FDC3 models (Figures 4, 5, 7, and 9). This is valuable for medical applications since it has been recently speculated that the lifetimes of the intermediates are crucial for recognizing amyloidogenic proteins.[51]

According to the FDC3 calculation the visible intermediates of Im7 are associated with states 2 and 4 (Figures 1 and 4). State 4 has a peak population of around 0.35. The folding process of Im9 follows a similar pathway where state 4 is now hardly detectable with a population around 0.015−0.050 both in the FDC and in the FDC3 scheme (Figures 1 and 5), in accord with the two-state nature of the folding mechanism. On the contrary, the mechanism of p16 initially takes two alternative paths that populate respectively state 2 or 5. The transient populations of the latter states eventually merge to give rise to the long-lived state 6 where IS helix B is simultaneously interacting with IS helices A and C (Figures 1 and 7). These features of state 6 are consistent with the picture of the transition state emerging from recent ϕ-value analysis of protein p16.[44] The existence of two alternative pathways of p16 eventually merging in state 6 agrees with the initial parallelism followed by different sequential steps that is expected to characterize funnel-like landscapes.[52] The long-lived state 4 of Im7 and state 6 of protein p16 meet the requisites for being considered gateway states[52] of the folding dynamics. Such gateway states are near native well populated, obligatory and rate-limiting conformations that are reached in the fast initial steps of folding. They correspond to the bottlenecks of the microroutes on the funnel-shaped energy surface.[52] These requisites are also met by states 2 and 5 respectively in the two-state folders Im9 and λ-repressor (Figures 1, 5, and 9).

The populations obtained within the FDC and the FDC3 models for Im7 and p16 exhibit interesting differences. State 4 of Im7 is approximately equally populated, both in the FDC and in the FDC3 scheme (Figure 4), but appears to be more persistent in the FDC3 plot. Also state 2 is present in both models but the FDC3 calculation predicts a larger peak population and a slower decay (Figure 4). The results obtained for state 4 of Im9 are approximately invariant with respect to the computational scheme adopted (Figure 5), save for a slightly larger peak population and a slower decay in the FDC3 model. For p16 the intermediate states 2, 5, and 6 are more visible in the FDC3 diagram than in the FDC plot (Figure 7). State 6 undergoes a much larger increase in population and decays at a sensibly slower rate in the FDC3 scheme than in the FDC calculation. Finally, state 5 dominates the folding process of the λ-repressor (Figure 9) and exhibits similar behavior in the FDC and FDC3 plots. The FDC3 calculation enhances moderately the populations of states 2 and 6 that were not visible in the FDC plot.

It is significant that our calculations for Im7 and p16 show that some intermediate states are much less visible and the folding time is grossly different from the experimental value when we abandon the FDC3 scheme and apply the bare FDC model. States 2 and 4 for Im7 in Figure 4 and state 6 for p16 in Figure 7 are two examples where the existence of long-lived intermediates is correctly reproduced only within the FDC3 scheme. This is a successful test of the ability of the FDC3 scheme to display the dominant intermediate state of folding that is proper to three-state folders. However, somewhat less neat results are obtained for two-state folders λ-repressor and Im9 since the FDC calculation shows that intermediates (in the sense of the partition of Figure 1) exist with nonnegligible population, which seems inconsistent with two-state folding mechanisms. This is essentially the case of state 2 for Im9 (Figure 5) and state 5 for the λ-repressor (Figure 9). Therefore, although the kinetic data on the folding times are satisfactory (Table 5) the detailed description of the folding pathway needs a more thorough discussion. In particular the most critical issue is to what extent the partition in states based on the aggregates of increasing order (Figure 1) reflects the presence of experimentally detectable intermediates.

Some considerations may help to deal with such a puzzling problem. Relevant to our discussion is the preliminary observation that the same protein can exhibit both two-state or three-state folding in response to varying environmental conditions.[34] However, such a level of detail is not attainable by the FDC/

FDC3 schemes that in the present version cannot discriminate between different environmental parameters. This is due to the very schematic description of the interactions between protein and solvent via the diffusion coefficient and temperature (see Methods and Theory). Analogous limitations of the FDC/FDC3 schemes, pertaining to the need of a more accurate representation of the intramolecular interactions. are discussed by M. Compiani et al. (in preparation). Therefore the theoretical predictions of the FDC/FDC3 models are not expected to capture these subtleties of the folding mechanism.

As far as the unexpected emergence of populated intermediates is concerned, it is to be noted that similar problems have been encountered on applying the DC scheme to two-state folders. The analysis carried out in ref 30 points out that the emergence of intermediates in DC models is quite sensitive to the values of the β_i^1s. This is the case of 1EBD C-chain and 1BDC where adjusting the estimates for β_i^1s enhances or alternatively depresses the population of an intermediate state, though these proteins are reported to have two-state folding. This indicates that possible inaccuracies in the determination of the β_i^1s may be conducive to spurious increase of the probability of occurrence of some intermediate states. Within the framework of the FDC/FDC3 models such an uncertainty is mainly due to the noise affecting the entropy signal[32] (see also final remarks in the present Discussion section).

Concerning the relationship between the intermediates envisaged by the partition of Figure 1 and the intermediate states that are experimentally detected, our view is that it appears more reliable to associate the experimental intermediates with the theoretical aggregates of higher rank $k > 2$. The 2-fold reason for these clusters to be more easily observable is that they do have a larger number of stabilizing interactions and also that interhelical contacts formed early during folding are expected to be more unstable than those formed later (as suggested by hydrogen exchange experiments[47]). Accordingly, states laying close to the unfolded state (see the graph of Figure 1) are less significant than those that are closer to the final state 8. This is the case of state 2 as compared with state 4 for Im7 and Im9, as well as states 2 and 5 in comparison with state 6 for p16. Consistent with this interpretation is the finding that state 4 is hardly visible in the FDC3 and FDC plots of Im9, which is known to have a two-state folding mechanism, whereas state 2 is well populated independently of the model used (see Figure 5). In a similar fashion, the FDC3 and FDC plots of the λ-repressor exhibit a negligible population for state 6 and more substantial probability for state 5, which represents the initial stage of the favorite folding pathway.

A final comment concerns the accuracy of the folding rates of Table 5. As a premise, we can state that quite crucial within our computational scheme is the reliability of the entropy signal that is used to determine the location of foldons, the length and location of the native helices, and also the stability of the IS helices. We have already mentioned that the noise affecting the entropy signal propagates down to the β_i^1s and affects also the population levels of the possible intermediates. Similarly, inaccuracies in the prediction of the native helices (particularly the position of their end residues) increases the uncertainty of the folding rates. The role of these parameters in determining the estimate of the folding rate can be appreciated in the Methods and Theory section. Somewhat analogous effects were investigated in ref 30 by changing systematically the length of interhelical loops.

A typical manifestation of the noise blurring the entropy signal is the defective prediction of helices 3 and 4 of the λ-repressor (see legend to Table 4 and Figure 8). In this case the neural network prediction incorporates helix 3 and helix 4 into a unique helical region. To optimize prediction we adopted a hybrid strategy in which we considered the crystallographic boundaries of helix 4 to determine the geometric properties of IS helix B, nevertheless taking into account that helix 3 had been detected by the neural predictor. This is achieved by renormalizing the loop length between IS helix A and B as detailed in the legend to Table 4.

V. Conclusion

In conclusion, the scope of the FDC model is enlarged as we have demonstrated that the FDC3 variant can effectively account for several features of the folding kinetics of three-state proteins. The general meaning for a theory of protein folding is that the same model describes successfully the kinetics of two-state and three-state folders, where the validity of previous criteria (e.g. contact order[48]) is limited to two-state proteins.[49] On the other hand, other empirical methods for the determination of the folding rate with broader validity[50] are based on features that are somehow incorporated in the FDC/FDC3 model. On the whole, such methods rely on generic parameters such as the contact order, protein length, and content in secondary structure. Clearly, a direct comparison is not possible since the FDC/FDC3 models do not depend explicitly on contact order or chain length. However, it is quite safe to state that the FDC/FDC3 models do not contradict the essence of these methods. In fact protein length is related to the maximal separation in sequence of the foldons whereas contact order is conceptually akin to the mean separation in sequence of the foldons, provided we consider the interhelical contacts among the residues in the foldons as the key contacts which determine the folding process. Data on the participation of these residues in the transition state provide preliminary evidence that this may be true (M. Compiani, E. Capriotti, and M. Vendruscolo, in preparation). These findings and the fact that foldons belong to nativelike secondary structures are also consistent with the extended nucleus theory,[18] which depicts the transition state as a distorted version of the native state. The successful application of the FDC3 model hints at the remarkable flexibility exhibited by the FDC scenario in describing folding processes with different degrees of complexity and cooperativity. As far as the general theory of protein folding is concerned, this gives credit to the view that the diffusional dynamics of foldons might be a promising basis for a unified theory of the folding of helical proteins.

Acknowledgment. This paper is based on the results obtained by A.S. in his thesis work (Stochastic models of folding of proteins with complex folding dynamics), discussed in the Department of Mathematics and Physics of the Catholic University of Brescia (Italy), under the supervision of M.C.. During the preparation of his thesis A.S. has benefited from the hospitality of Prof. Rita Casadio at the Biocomputing Unit of the Centro Interdipartimentale di Ricerche Biotecnologiche, University of Bologna (Italy). This work has been partially funded by the Italian Ministry for Research with a PRIN 2002 grant delivered to M.C. E.C. acknowledges financial support from the CNR/MURST project "Development and implementation of algorithms for predicting protein structure", the CNR project "Molecular genetics and functional genomics", and a FISR2002 project.

References and Notes

(1) Baldwin, R. L. *Nature Struct. Biol.* **1999**, *6*, 814−817.

(2) Casadio, R.; Compiani, M.; Fariselli, P.; Jacoboni, I.; Martelli, P. L. *SAR QSAR Environ. Res.* **2000**, *11*, 149−182.
(3) Sali, A. *Nature Struct. Biol.* **1998**, *5*, 1029−1032.
(4) Kelly, J. W. *Curr. Opin. Struct. Biol.* **1996**, *6*, 11−17.
(5) Dobson, C. M. *Trends Biol. Sci.* **1999**, *24*, 329−332.
(6) Bousset, L.; Thomson, N. H.; Radford; S. E.; Melki, R. *EMBO J.* **2002**, *21*, 2903−2911.
(7) Guerois, R.; Serrano, L. *Curr. Opin. Struct. Biol.* **2001**, *11*, 101−106.
(8) Thirumalai, D.; Woodson, S. A. *Acc. Chem. Res.* **1996**, *29*, 433−439.
(9) Dill, K. A.; Chan, H. S. *Nature Struct. Biol.* **1999**, *4*, 10−19.
(10) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins*: *Struct. Funct. Genet.* **1995**, *21*, 167−195.
(11) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−744.
(12) Zwanzig, R. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *89*, 20−22.
(13) Compiani, M.; Fariselli, P.; Martelli, P. L.; Casadio, R. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9290−9294.
(14) Panchenko, A. R.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 2008−2013.
(15) Yu, M.-H.; King, J. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 6584−6588.
(16) Panchenko, A.; Luthey-Schulten, A.; Cole, R.; Wolynes, P. G. *J. Mol. Biol.* **1997**, *272*, 95−105.
(17) Compiani, M.; Capriotti, E.; Casadio, R. *Phys. Rev. E*: *Stat. Phys. Plasmas, Fluids, Relat. Interdiscip. Top.* **2004**, *69*, 051905(1−8).
(18) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1525−1529.
(19) Kim, P. S.; Baldwin, R. L. *Annu. Rev. Biochem.* **1982**, *51*, 459−489.
(20) Baldwin, R. L.; Rose, G. D. *Trends Biochem. Sci.* **1999**, *24*, 26−33; 77−83.
(21) Serra, R.; Zanarini, G.; Andretta, M.; Compiani, M. *Introduction to the Physics of Complex Systems*; Pergamon Books: Oxford, UK, 1986.
(22) Rumbley, J.; Hoang, L.; Mayne, L.; Englander, S. W. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 105−112.
(23) Baldwin, R. L. *Nature Struct. Biol.* **2001**, *8*, 92−94.
(24) Karplus, M.; Weaver, D. L. *Nature* **1976**, *260*, 404−406.
(25) Karplus, M.; Weaver, D. L. *Biopolymers* **1979**, *18*, 1421−1437.
(26) Karplus, M.; Weaver, D. L. *Protein Sci.* **1994**, *3*, 650−668.
(27) Yapa, K. K.; Weaver, D. L. *J. Phys. Chem.* **1996**, *100*, 2498−2509.
(28) Pappu, R. V.; Weaver, D. L. *Protein Sci.* **1998**, *7*, 480−490.
(29) Burton, R. E.; Myers, J. K.; Oas, T. G. *Biochemistry* **1998**, *37*, 5337−5343.
(30) Islam, S. A.; Karplus, M.; Weaver, D. L. *J. Mol. Biol.* **2002**, *318*, 199−215.
(31) Freire, E.; Murphy, K. P.; Sanchez-Ruiz, J. M.; Galisteo, M. L.; Privalov, L. P. *Biochemistry* **1992**, *31*, 250−256.
(32) Compiani, M.; Fariselli, P.; Casadio, R. *Phys. Rev. E*: *Stat. Phys. Plasmas, Fluids, Relat. Interdiscip. Top.* **1997**, *55*, 7334−7343.
(33) Capaldi, A. P.; Kleanthous, C.; Radford, S. E. *Nature Struct. Biol.* **2002**, *9*, 209−216.
(34) Ferguson, N.; Capaldi, A. P.; James, C.; Kleanthous, C.; Radford, S. E. *J. Mol. Biol.* **2002**, *286*, 1597−1608.
(35) Gianni, S.; Guydosh, N. R.; Khan, F.; Caldas, T. D.; Mayor, U.; White, G. W. N.; DeMarco, M. L.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13286−13291.
(36) Ptitsyn, O. B.; Ting, K.-L. H *J. Mol. Biol.* **1999**, *291*, 671−682.
(37) Martinez, J. C.; Serrano, L. *Nature Struct. Biol.* **1999**, *6*, 1010−1015.
(38) Mirny, L.; Shakhnovich, E. I. *J. Mol. Biol.* **1999**, *291*, 177−196.
(39) Plaxco, K. W.; Larson, S.; Ruczinsky, I.; Riddle, D. S.; Thayer, E. C.; Buchwitz, B.; Davidson, A. R.; Baker, D. *J. Mol. Biol.* **2000**, *298*, 303−312.
(40) Casadio, R.; Compiani, M.; Fariselli, P.; Martelli, P. L. *ISMB* **1999**, *7*, 66−76.
(41) Bashford, D.; Weaver, D. L.; Karplus, M. *J. Biomol. Struct. Dyn.* **1984**, *1*, 1243−1255.
(42) Tang, K. S.; Guralnick, B. J.; Wang, W. K.; Fersht, A. R.; Itzhaki, L. S. *J. Mol. Biol.* **1999**, *285*, 1869−1886.
(43) Zhang, P.; Peng, Z. *J. Mol. Biol.* **2000**, *299*, 1121−1132.
(44) Tang, K. S.; Fersht, A. R.; Itzhaki, L. S. *Structure* **2003**, *11*, 67−73.
(45) Jackson, S. E. *Folding Des.* **1998**, *3*, R81−R91.
(46) Freire, E.; Murphy, K. P. *J. Mol. Biol.* **1991**, *222*, 687−698.
(47) Jennings, P. A.; Wright, P. E. *Science* **1993**, *262*, 892−896.
(48) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985−994.
(49) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. *Proteins Struct. Funct. Gen.* **2003**, *51*, 162−166.
(50) Ivankov, D. N.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 8942−8944.
(51) Ramirez-Alvarado, M.; Merkel, J. S.; Regan, L. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8979−8984.
(52) Schonbrun, J.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12678−12682.