Dynamics of the minimally frustrated helices determine the hierarchical folding of small helical proteins

M. Compiani,[1,2] E. Capriotti,[2,3] and R. Casadio[2,4]
[1]*Department of Chemical Sciences, University of Camerino, Camerino, Italy*
[2]*CIRB, Centro Interdipartimentale per la Ricerca Biotecnologica, University of Bologna, Bologna, Italy*
[3]*Department of Physics, University of Bologna, Bologna, Italy*
[4]*Department of Biology, University of Bologna, Bologna, Italy*

(Received 3 October 2003; revised manuscript received 9 December 2003; published 12 May 2004)

In this paper we aim at determining the key residues of small helical proteins in order to build up reduced models of the folding dynamics. We start by arguing that the folding process can be dissected into concurrent fast and slow dynamics. The fast events are the quasiautonomous coil-to-helix transitions occurring in the minimally frustrated initiation sites of folding in the early stages of the process. The slow processes consist in the docking of the fluctuating helices formed in these critical sites. We show that a neural network devised to predict native secondary structures from sequence can be used to estimate the probabilities of formation of these helical traits as they are embedded in the protein. The resulting probabilities are shown to correlate well with the experimental helicities measured in the same isolated peptides. The relevance of this finding to the hierarchical character of folding is confirmed within the framework of a diffusion-collision-like mechanism. We demonstrate that thermodynamic and topological features of these critical helices allow accurate estimation of the folding times of five proteins that have been kinetically studied. This suggests that these critical helices determine the fundamental events of the whole folding process. A remarkable feature of our model is that not all of the native helices are eligible as critical helices, whereas the whole set of the native helices has been used so far in other reconstructions of the folding mechanism. This stresses that the minimally frustrated helices of these helical proteins comprise the minimal set of determinants of the folding process.

DOI: 10.1103/PhysRevE.69.051905 PACS number(s): 87.15.Cc, 87.14.Ee

Recently it has been argued that the whole gamut of possible folding mechanisms spans the range between two extreme models. They are the diffusion-collision model (DC model) and the nucleation-condensation scenario (NC model) [1,2] and emphasize, respectively, the hierarchical and the cooperative component of folding. It has also become clear that managing the complexity of protein folding requires simplification strategies that rely on minimalist models of the dynamical processes involved [3–7]. Subscribing to a hierarchical view of folding, one is allowed to exploit the modularity of the folding process to uncouple the formation of global and local structures [1,8,9]. Within this approach, models depicting the formation of simple elements of the secondary structure are intended to shed light on the dynamics and kinetics of elementary events of folding. Numerous theoretical studies, pioneered by the Zimm-Bragg theory (ZB theory) [10], have been devoted to helix-coil transitions viewed as the simplest stages of folding. In the same vein, new theoretical and experimental approaches to β-hairpin formation have been recently proposed [11–13]. Renewed interest in the processes of helix formation is due to their importance in the context of bottom-up strategies for the rational design of proteins [14,15] as well as in hierarchical models of folding [1,8,9,16], in which stabilization of elements of secondary structure precedes the formation of tertiary interactions. The prototype of the quantitative hierarchical theories of folding is the DC model [17–19]. It depicts the folding of proteins in terms of stochastic encounters of marginally stable microdomains which, in the case of helical proteins, coincide with the native helices.

Hints as to the modularity of the folding dynamics of helical proteins are to be found in Ref. [20] where we have argued that dissection of folding in temporally distinct events is feasible in the minimally frustrated initiation sites of folding (ISs) that trigger the nucleation of the native IS-containing helices (NIS helices). As in Ref. [20], we restrict the present analysis to helical proteins in which the search for the ISs is especially successful.

In the present paper we pursue the goal to identify the minimal set of determinants of the folding dynamics and present an effective tool for the calculation of the folding rates that is based mainly on sequence information. To this aim, we first set about demonstrating the existence of helical building blocks that are formed in the course of fast elementary helix-coil transitions taking place in the ISs. These processes result in the stabilization of the precursors of the NIS helices (IS helices, for brevity). The IS helices are transiently stabilized in the early stages of folding, i.e., prior to the establishment of any appreciable amount of tertiary structure, when short-range interactions (acting between residues that are close in sequence) overcome the long-range forces (acting between residues that are distant in sequence). More precisely, we show that the average thermodynamic properties that characterize the formation of the IS helices within the whole protein, are comparable with those of the same peptides when they are excised from the protein. Second, the relevance of the IS helices to the folding mechanism is definitely demonstrated as we show that this set of helices is sufficient to reconstruct the folding dynamics. To do this we hypothesize that the IS helices are the fluctuating micro-

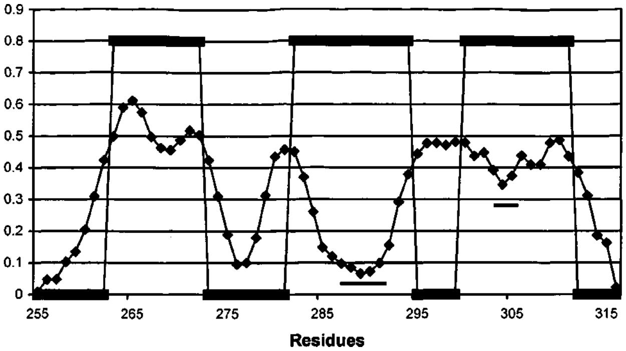

FIG. 1. Information entropy profile vs residue number of the crystallized fragment (255–316) of thermolysine (PDB code, 1TRL) calculated according to the procedure of Ref. [20] (see text). The step function superimposed to the curve indicates the location of the α-helices predicted by the neural network. The ISs of folding are defined as the segments classified as helical regions by the neural network that correspond to the minima of the entropy plot that lie below a threshold entropy $S=0.416$ [20,21]. The ISs comprise those residues that deviate from the entropy minimum by less than 0.05 [20]. The predicted helices 282–294 and 300–311 contain two minima that fulfill this criterion. The ISs span the 287–291 and 303–306 regions (black bars). The first helix does not include any IS.

domains that participate in the folding mechanism depicted as a comparatively slow diffusion-collision process. Using the DC model [17–19] and the thermodynamic characterization of the IS helices we eventually get accurate predictions of the folding times of small helical proteins with two-state folding. Finally, we discuss the relationships between our model and the DC and NC pictures of protein folding.

I. ESTIMATING THE STABILITY OF THE IS HELICES

As a first step we identify the ISs in the sequence of the protein under study. We have shown that the output of a neural network can be processed in order to identify the putative regions of the sequence corresponding to the ISs of folding of helical proteins [20,21]. For the nth residue being classified, the output vector can be viewed as a discrete probability function $[p_h(n), p_c(n)=1-p_h(n)]$ defined in the space of structures [α-helix (h) and random coil (c)]. $p_h(n)$ is the probability that the residue in position n has helical structure. The Shannon information entropy for the nth residue $S(n)=-p_h(n)\ln p_h(n)-[1-p_h(n)]\ln[1-p_h(n)]$ has been shown to be equivalent to a measure of frustration [22], and can be used to draw an entropy profile along the protein (Fig. 1). We have proved that the ISs are found in the lowest entropy minima corresponding to helical structures [20]. Note that the very definition of an IS implies that its residues trigger the nucleation step of the coil-to-helix transition of the IS helices they are in. According to the ZB theory [10], nucleation is followed by the elongation step which results in the formation of the IS helical structure.

To address the thermodynamics of the IS helices we derive their energy landscape from the entropy profile. To elucidate the relationship between the thermodynamics of the coil-to-helix transition for any residue and the network's output, let us depict the growth of the helix by means of a free energy profile, $\Delta g(n)=g(n)^{helix}-g(n)^{coil}$, specifying the free energy cost for the transition of the residue in position n [10,23]. Within a continous picture it is straightforward to check that the relationship between $S(n)$ and the free energy profile is $\Delta g(n)=S'(n)[p_h'(n)]^{-1}$ (primes indicate derivatives with respect to n).

The classical scheme for passing from the single-residue thermodynamics to the thermodynamics of the whole helix is the ZB theory [10] that is suited to deal with homopolymers having constant length. However, natural peptides usually have nonuniform compositions. In addition, they are commonly viewed as flickering elements undergoing fluctuations in length on quite short time scales, as compared with the time scale of the measuring apparatus or the typical times of formation of tertiary contacts in protein folding [24–28]. This is consistent with the dissection of the folding process in fast and slow dynamical subprocesses. Borrowing the terminology of the DC model, the IS helices can be viewed as fluctuating microdomains that, in the early steps of folding and under the action of short-range interactions, reach a temporary equilibrium conformation that concludes the local fast dynamics. In the meantime, the IS helices undergo a slow diffusion-collision process that mimics the global dynamics. Such a clear-cut difference in the time scales suggests that the simplest way to generalize the ZB theory into a description of the fluctuating IS helices is to devise an equivalent ZB model that is averaged over the possible lengths of the helix at issue. It should be also remarked that relying on the neural network to implement the averaged ZB description automatically takes into account the chemical heterogeneity of the protein segment.

Using the ZB formalism, we express the equilibrium constant K_n for the formation of an n-residue helix in terms of the equilibrium constants for nucleation ($K_n^{nucl}=\sigma s$) and elongation ($K_n^{elong}=s^{n-1}$) as $K_n=\sigma s^n$. Moreover, we take into account that helices fluctuate among different folded states with a variable number, $i\in[1,n]$, of helical residues. To make contact with the experimental helicity measurements, we assume that the effective equilibrium constant K_{eff} corresponding to the measured helical contents is approximated by the average constant

$$K_{eff}=\sum_{i=1}^{n} K_i/n=\sum_{i=1}^{n} \sigma s^i/n. \qquad (1)$$

Setting $\sigma s^i=\exp(-\Delta G_i)$, where ΔG_i is the free energy of formation (in RT units) of an i-residue helix, we can rewrite Eq. (1) as

$$K_{eff}=[\exp(-\Delta G_1)-\exp(-\Delta G_n)]/[n(1-s)]. \qquad (2)$$

Within a continuous picture, the total free energy ΔG_i is written as ΔG_x (as a function of a continuos coordinate x) and can be calculated as $\Delta G_x=\int_{x_0}^{x}\Delta g(y)dy$, where x_0 represents the coordinate of the local entropy minimum, where the nucleation process is expected to occur. Using the second theorem of the mean [29] to compute the jump of the function $\exp(-\Delta G_x)$ between the first and the nth residues (in position x_0 and x, respectively), Eq. (2) becomes

$$K_{eff}=\exp\left[-\int_{x_0}^{\xi}\Delta g(x)dx\right]\Delta g(\xi)/(1-s), \qquad (3)$$

where $\xi\in[x_0,x]$. The exponential in Eq. (3) can be transformed by using the above mentioned relationship between $\Delta g(x)$ and $S(x)$. Then integrating by parts and applying the second theorem of the mean [29] to the resulting integral leads to

$$\exp\left[-\int_{x_0}^{\xi}\Delta g(x)dx\right]=\exp\{-[S(\xi)-S(x_0)][p_h'(z)]^{-1}\}, \qquad (4)$$

where $z\in[x_0,\xi]$. The final form of Eq. (3) as a function of entropy is

$$K_{eff}=\exp\{-[S(\xi)-S(x_0)]/[p_h'(z)]\}S'(\xi)[p_h'(\xi)]^{-1}(1-s)^{-1}. \qquad (5)$$

A more manageable form of Eq. (5) ensues if we suppose that $S(x_0)$ is negligible compared to $S(\xi)$. This condition is certainly fulfilled by good minima that have very low values of information entropy. Moreover, the last two factors in Eq. (5) provide logarithmic corrections to the exponent of the exponential term. Their contribution can be included in the integral, provided we shift the upper limit of integration ξ of Eq. (4) to a suitable value y (correspondingly $z\in[x_0,y]$). For example, the correction due to $(1-s)^{-1}$ can be readily evaluated provided that the physically meaningful condition $s<1$ is met. In this case the correction translates to a shift of ξ towards x_0. On expanding the exponential to first order, Eq. (5) takes the form

$$K_{eff}=\{1-S(y)/p_h'(z)\}S'(y). \qquad (6)$$

Retaining the factorization of K_{eff} in the product of an effective nucleation constant and an effective elongation constant ($K_{eff}=K_{nucl}K_{elong}$), we identify the factor in braces in Eq. (6) with K_{nucl} and factor $S'(y)$ with K_{elong}. For lack of explicit estimates for y and z we devise an approximated graphical procedure for evaluating K_{nucl} and K_{elong} from the entropy profile. On account of the general structure of Eq. (6), we set $K_{nucl}\approx 1-S_{min}$, where S_{min} is related to the entropy minimum. We use the simplest choice $S_{min}=S(x_0)$ to evaluate the helicities of the IS helices. A slightly different choice has proven more effective in the calculation of the folding rates, as illustrated below and in the legend of Fig. 2. For the elongation constant we have set $K_{elong}\approx S'_{\min}$, where $S'_{\min}=\min\{\boldsymbol{\nabla}^L S,\boldsymbol{\nabla}^R S\}$ and $\boldsymbol{\nabla}^L S$ and $\boldsymbol{\nabla}^R S$ are the average entropy gradients in the left and right sequence regions flanking the IS (see Fig. 2).

The constant K_{eff} lends itself to the calculation of the helicity β, i.e., the probability that a peptide is in the helical state. In order to compute β for an n-residue helix we recall that, in the case of strong cooperativity ($\sigma\ll 1$), $\beta\approx K_n=K_n^{nucl}K_n^{elong}$ (see also Ref. [31]). Thus, using the approximation $\beta\approx K_{eff}$ for a fluctuating IS helix, we can estimate β by means of the graphical estimation devised in Fig. 2.

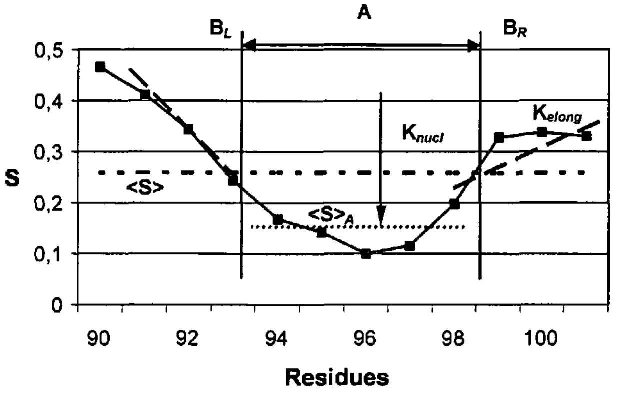

FIG. 2. Derivation of thermodynamic parameters of the IS helices from the information entropy profile, through a graphical estimation of Eq. (6). The curve reproduces a stretch of the entropy plot corresponding to the last NIS helix of protein 1HRC (see Table I and Fig. 4). The entropy minimum $S(x_0)$ is found in $x_0=96$. The horizontal dashed line signals the average entropy $\langle S\rangle$ of the NIS helix under study. The residues are divided in two subsets $\mathcal{A}=\{\text{residues with } S<\langle S\rangle\}$ and $\mathcal{B}=\{\text{residues with } S\geqslant\langle S\rangle\}=\mathcal{B}_L\cup\mathcal{B}_R$. The horizontal dotted line represents the average entropy $\langle S\rangle_A$ calculated over the set $\mathcal{A}$. The vertical arrow represents $1-\langle S\rangle_A$ that estimates K_{nucl}. This amounts to setting $S_{min}=\langle S\rangle_A$ in the graphical estimate of $K_{nucl}\approx 1-S_{min}$ referred to in the discussion to Eq. (6). This choice has given optimal results in implementing the DC model. As explained in the text, the alternative choice $S_{min}=S(x_0)=S(96)$ made on comparing CD data and calculated helicities (see Fig. 3), implies that the set $\mathcal{A}$ has shrunk to x_0. To compute the average slopes $\boldsymbol{\nabla}^L S$ and $\boldsymbol{\nabla}^R S$ of the entropy profile (dashed lines in $\mathcal{B}_L$ and $\mathcal{B}_R$), a least squares procedure has been applied to the data in $\mathcal{B}_L$ and $\mathcal{B}_R$.

II. THE IS HELICES AS BUILDING BLOCKS OF FOLDING

Having computed the helicity β we are in a position to investigate the role of the IS helices as building blocks in a modular mechanism of protein folding. To this aim we inquire whether the thermodynamic features of any IS helix within the native protein can be extrapolated to the same isolated peptide. To perform this test we have taken from the literature a set of isolated peptides comprising an IS of the parent protein and for which the helicities have been determined experimentally through circular dichroism (CD) measurements [33] (see legend of Fig. 3). The β value of each IS helix has been calculated from the related entropy profile following the procedure shown in Fig. 2, and then has been compared with the corresponding experimental CD value (Fig. 3).

We do not expect a perfect correspondence between the absolute values of CD and calculated helicities (β). The main reason is that the thermodynamic properties of the IS helices have been computed using the native entropy landscape of the NIS helices derived from the native structure of the protein. This procedure would closely approximate the experimental results in the case that the tertiary interactions give a

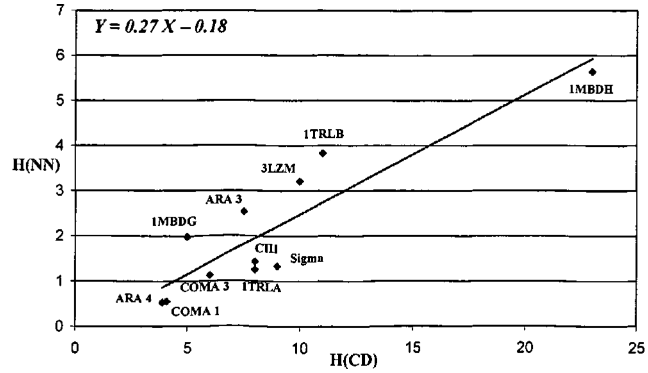

FIG. 3. Comparison of the helicities $H(NN)$ calculated through the neural network (NN) and experimental helicities $H(CD)$. Circular dichroism values (CD) were measured on the isolated peptides indicated in the panel. The labels for ARA, COMA, CIII, 3LZM, and Sigma are the same used in the original paper reporting the CD values (see Table 3 of Ref. [34]). In addition we have considered the data for two helical peptides of the thermolysine segment (shown in Fig. 1) (spanning the 281–295 and the 301–310 regions), that were drawn from Ref. [35], and the H and G helices of myoglobin (PDB code, 1MBD), that were studied in Ref. [36]. The NN values were estimated by processing the full sequence of the above mentioned proteins with our neural network-based procedure. The experimental helicities $H(CD)$ are usually expressed as mean helicity of each of the N residues of the peptide, where N is the number of the helical residues as determined from crystallographic data. To remedy the discrepancies between N and the number of helical residues predicted by the neural network, n, we set $H(CD)N=n\beta$. The scatter plot compares $H(CD)$ with $H(NN)=\beta n/N$ ($H(CD)$ and $H(NN)$ have been expressed in percent). The equation in the left corner refers to the least square line.

nearly vanishing contribution to the native structure. However, we must bear in mind that although the NIS helices are minimally affected (compared to the rest of the protein) by the long-range interactions that intervene in the late stages of folding, the effects of these forces are not completely negligible even in these minimally frustrated regions of the protein. This introduces a deviation between the properties of the isolated peptides (not subjected to any tertiary interactions) and the native entropy landscape used to evaluate the β's of the IS helices. A further source of discrepancies can be traced back to the approximations inherent in Eq. (6) and its graphical estimation (Fig. 2). The issue of the level of noise affecting the entropy profile has been addressed in Ref. [22].

Nonetheless, if the NIS helices preserve to some extent the substantial independence on the context that is typical of the ISs and the IS helices, we are confident that some relationship exists between the β values and the CD helicities. Actually, the two sets of values shown in Fig. 3 turn out to be satisfactorily correlated with correlation $\rho=0.9$. This indicates that the essential factors determining the helicity in the isolated peptides seemingly determine the trend of the helical content of the IS helices in the protein's interior. It ensues that even in the context of the whole protein the IS helices exhibit the character of semi-independent elements. Such a minimization of the conflicts between local and global interactions [37] allows us to conclude that, in a sense, the minimally frustrated character of the ISs [20] is inherited by the IS helices.

Further evidence that the formation of each IS helix within the full protein parallels, in the average sense described above, the process occurring in the isolated peptide, is provided by the simulations carried out on the isolated helices of apomyoglobin [38]. The computational data and our ranking of the IS helices according to the β values agree in pointing out the largest stability of the NIS helices G and H, as compared to the other helical regions.

According to the ZB picture, the coil-to-helix transition in the nucleation site is expected to be faster than the comparatively slow elongation process. Therefore, kinetic studies can be useful to probe our contention that the putative nucleation site of an IS helical segment corresponds to the minimum of the entropy plot. Time-resolved experiments [39] on the helical segments of apomyoglobin indicate that the central regions of helices G and H (corresponding to segments 102–115 and 133–143, respectively) undergo the fastest transitions to the helical conformation. Therefore, they are likely to include the nucleation sites of the G and H helices. This is fully confirmed by our predictions, since the entropy criterion [20] identifies the IS of the IS helix G with the 109–116 segment, and the IS of the IS helix H, with the 125–143 region.

A further test can be carried out on helix 4 of cytochrome c in which the essential steps of formation have been monitored with NMR experiments [40]. The temporal ordering of the events as revealed experimentally indicates that residue 95 acquires its helical conformation in the shortest time, followed by residue 91 and eventually by residue 100. This finding is fully confirmed by our analysis. The entropy plot clearly exhibits a deep minimum at residue 96, whereas the IS spans the 95–97 segment (see Fig. 4).

III. RECONSTRUCTION OF THE FOLDING DYNAMICS

In order to provide more cogent evidence as to the fundamental role of the IS helices in the folding mechanism, we show that these same helical regions can be used to reconstruct the whole folding process. As a matter of fact we demonstrate that the IS helices provide the minimal set of determinants which rule the full kinetics of the folding process. The test consists in computing the folding times of helical proteins with the DC model, where we equate the IS helices with the microdomains undergoing the diffusion-collision processes. The small helical proteins used as a benchmark and the corresponding IS helices are listed in Table I. The implementation of the DC model follows the standard version [17–19,30–32].

The characteristic time τ_{ij} for coalescence of the colliding microdomains (labeled i and j) can be evaluated as

$$\tau_{ij}=\frac{G}{D}+\frac{VL(1-\beta_{ij})}{AD\beta_{ij}}. \qquad (7)$$

In Eq. (7) A is the sum of the areas of the colliding microdomains that in the spherical approximation are ascribed the radii R_i and R_j. D is the relative diffusion coefficient defined

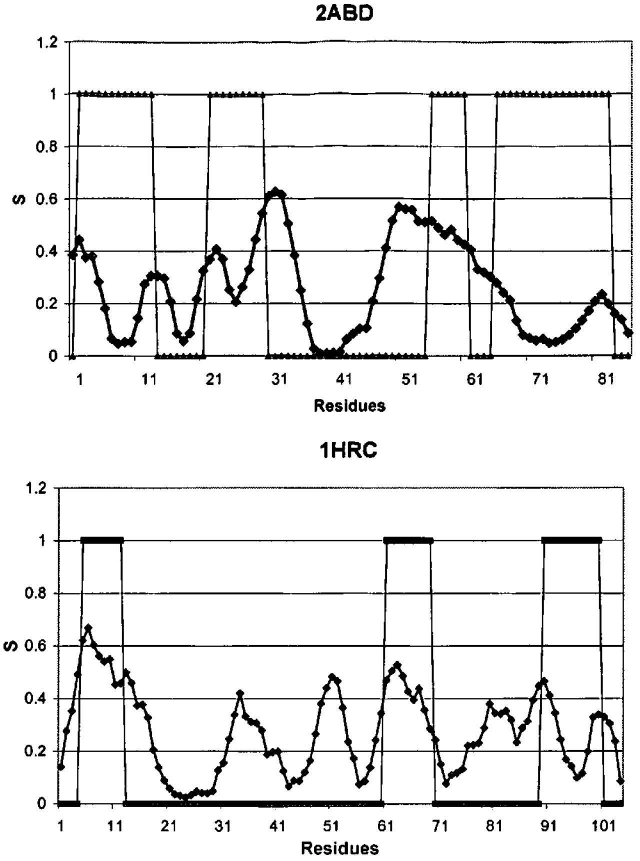

FIG. 4. Entropy profiles of 2ABD and 1HRC, based on the secondary structure prediction of the neural network. The average entropy of the ISs of 1HRC, estimated at the minima of the 60s and 90s NIS helices (see Table I) is clearly higher than the average level of the entropy minima of the NIS helices 1,2, and 4 of 2ABD. The effects of noise in the entropy signal are visible in that two very short helices are not predicted by the neural network.

as $D=k_BT(R_i^{-1}+R_j^{-1})/6\pi\eta$, where η is the viscosity. G and L are defined as

$$G=-\frac{R_{max}^2(5-9\epsilon+5\epsilon^3-\epsilon^6)}{15\epsilon(1-\epsilon^3)}, \qquad (8)$$

$$L^{-1}=\frac{1}{R_{min}}+\alpha\frac{\alpha R_{max}\tanh[\alpha(R_{max}-R_{min})]-1}{\alpha R_{max}-\tanh[\alpha(R_{max}-R_{min})]}. \qquad (9)$$

G and L depend on the geometric parameters $R_{min}=R_i+R_j$, $R_{max}=R_{min}+$linker length and on $\alpha=(D\tau)^{-1/2}$, $\epsilon=R_{min}/R_{max}$ and $V=4\pi(R_{max}^3-R_{min}^3)/3$. τ is the time constant for the coil-helix transition for which we have assigned the values 0.1 and 1 ns (Table II). These values are consistent with the values currently used and suggested in the literature [31,41–43]. The choice of the appropriate τ has been made according to the following heuristic criterion: the lower the entropy of the IS, the more biased the configurational probability density (estimated by the output of the neural network) towards the helical structure. This, in turn, implies that the residues at hand (belonging to any foldon) are more prone to reside in helical configurations. Consequently, the search time to reach the helical state is shorter. In a refined version of the present diffusion-collision calculation one can tune the τ parameter for each IS helix. Here we have made use of a more rough criterion, where we have taken into account the average entropy value of all the ISs directly from the entropy plot of the protein under study. In Fig. 4 the entropy profiles of 2ABD and 1HRC clearly exemplify the difference between a protein endowed with low-entropy IS helices (2ABD) and a protein with high-entropy IS helices (1HRC). Accordingly, we use a small τ (0.1 ns) for proteins with helices that are characterized by low information entropies, and a comparatively larger τ (1 ns) for proteins with ISs that are characterized by higher information entropies.

Predicting the structure and the ISs of small proteins like 1ENH is a tough problem for the neural network due to the small size of the protein. In this case it is likely that boundary effects have a more sensible influence on the reliability of the prediction. Such effects arise since the sliding input window of the network (17 residue long) has blank sites until the center of the window reaches the ninth residue (at the N-terminal), and as long as the $(N-9)$th residue (at the C-terminal) is trespassed. The entropy profile of 1ENH is noisy and less reliable in the C-terminal region of the sequence. To circumvent this difficulty we have used a single-sequence input for the neural network instead of the multiple-sequence alignments that have been adopted for the other four proteins. By so doing helix 3 of 1ENH turns out to be a NIS helix, although the entropy signal is still quite noisy.

In the standard DC model [31] the key parameter is the probability P_{ij} that takes into account the prerequisites for an effective collision to occur, i.e., a collision that results in the stable aggregation of the colliding microdomains, labeled i and j. When ineffective collisions occur (with probability $1-P_{ij}$) the microdomains bounce back and the diffusional dynamics starts anew. According to Ref. [31] we define P_{ij} as $P_{ij}^l=\gamma_i^n\gamma_j^m\beta_i^n\beta_j^m(l=m+n)$. Indexes $n,m\geqslant 1$ refer to the rank, i.e., the number of IS helices forming each microdomain. Clearly, $l\geqslant 2$.

The factor β_i^1 (folding probability) measures the probability that the structures of the ith IS helix is sufficiently close to the native form [31,32]. Estimates for β_i^1 are obtained according to the graphical procedure introduced above.

The factor γ_i^n (orientational probability) gives the probability that each microdomain approaches its partner with a favorable orientation. Estimates of γ_k^n in terms of geometrical features usually result from computing the ratio of the lost solvent accessible area to the total accessible area [31] (see Table I). Van der Waals volumes of the helices have been computed by means of TINKER; the program DSSP provided the accessible surfaces of the various helices as well as the surface that is lost upon contact. The parameters used in our calculations are collected in Tables I and II. Following the original implementations of the DC model we set $\beta_i^m=1$ for the aggregates with rank $m\geqslant 2$ [19,30–32]. This amounts to assuming that stable interactions are established between microdomains undergoing effective collisions. The

TABLE I. Folding probabilities and geometric features of the IS helices belonging to the five proteins examined in this paper. 1ENH is the engrailed homeodomain; 1LMB4 is the fourth chain of the λ repressor; 1IMQ is the Colicin E9 Immunity Protein Im9; 1HRC is the horse heart cytochrome *c* and 2ABD is the acyl-coenzyme A binding protein. The helices numbering in the second column is as usual from the *N* terminal to the *C* terminal. The residues in the third column correspond to the ends of the NIS helices predicted by the neural network. The IS helix column lists the range spanned by each IS helix. β_i^1 specifies the folding probability of the *i*th IS helix. The remaining three columns give the solvent accessible area, the volume of the native helices and the total accessible surface lost upon pairwise collision. The latter figure is expressed as the sum of the surfaces lost by each microdomain.

Protein	Helix	Residues	IS-helix	β_i^1	Area (Å^2)	Volume (Å^3)	Surface loss (Å^2)
1ENH	1	8–18	10–14	0.076	1610	1869	Pair(1–2): 244+234
	2	23–36	27–28	0.031	1828	2413	Pair(1–3): 197+184
	3	40–48	40	0.006	1294	1582	Pair(2–3): 321+286
1LMB4	1	9–26	13–22	0.099	2307	2467	Pair(1–4): 151+157
	4	59–70	63–66	0.085	1378	1445	Pair(1–5): 67+76
	5	79–90	82–83	0.048	1475	1578	Pair(4–5): 115+99
1IMQ	1	11–24	15–20	0.087	1544	1664	Pair(1–2): 289+267
	2	30–41	36–38	0.019	1725	1575	Pair(1–3): 224+245
	3	65–77	70–71	0.008	1505	1549	Pair(2–3): 76+59
1HRC	2	61–69	65–67	0.015	1388	1227	Pair(2–3): 309+330
	3	90–100	95–97	0.059	1672	1516	
2ABD	1	2–13	7–10	0.056	1397	1468	Pair(1–2): 62+60
	2	22–30	25–26	0.0005	1289	1241	Pair(1–4): 341+362
	4	66–83	70–77	0.018	2077	2385	Pair(2–4): 300+264

time evolution of the probabilities of the different aggregates is ruled by a master equation in which the transition probabilities per unit time are computed as $(\tau_{ij})^{-1}$ [31]. Following Ref. [31] 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 aggregate.

The DC-like dynamics of the IS helices has been used to compute the folding times of a set of five proteins whose kinetics has been experimentally investigated. The results are displayed in Table II.

TABLE II. Global parameters and comparison of the computed and experimental folding times of the five proteins listed in Table I. The first column indicates the PDB code of the proteins. *N* is the number of microdomains (i.e., IS helices) used in the FDC model. n_h is the number of native helices of each protein (i.e., the number of microdomains used in the standard implementation of the DC model). *L* is the length of the protein (total number of residues). τ is the adjustable relaxation time of the coil-helix transitions. The guiding principle for the choice of a short or a long τ is illustrated in the text. τ_{comp} denotes the computed folding time and τ_{exp} is the experimental folding time. The experimental times of 1ENH has been taken from Ref. [2]; the data for 1LMB4, 1IMQ, 1HRC, and 2ABD have been drawn from Ref. [50].

Protein	N	n_h	L	τ(ns)	$\tau_{comp}(\mu\text{s})$	$\tau_{\text{exp}}(\mu\text{s})$
1ENH	3	3	54	0.1	30	27
1LMB4	3	5	92	0.1	213	204
1IMQ	3	4	86	1	680	670
1HRC	2	5	105	1	2300	2500
2ABD	3	4	86	0.1	4400	5000

IV. DISCUSSION

Apparently, comparison with experimental data shows that our estimates of the folding times are fairly accurate (Table II). This indicates that our model has taken into account all the crucial events that determine the folding kinetics. These results are particularly intriguing since only a subset of the helical regions of each protein have been included in our simulations of the folding process ($N \leqslant n_h$, in Table II). In this respect the identification of the IS helices with the microdomains of the DC model introduces a meaningful novelty as compared with previous implementations of the DC dynamics in which all of the native helices of the protein under study were considered in the computational scheme [17–19,31,32,41,44,51].

The fact that despite this we capture the essentials of the folding dynamics, suggests that besides driving the initial steps of folding [20], the IS helices critically control the kinetics of the whole folding process. As a corollary, this implies that the settling of non-IS helices into their native structure occurs in the late stages of folding and is non-rate-limiting. Notwithstanding the different definition proposed in this paper, these characteristic features make the IS helices conceptually similar to the foldons that have been introduced in Refs. [45,46] as the ultimate determinants of folding. Therefore, we feel entitled to refer to our picture of folding as the foldon diffusion-collision model (FDC model).

Clearly, the prominent role assigned in the FDC model to the IS helices is consistent with the finding that preorganized elements with nativelike secondary structure play a major role in the overall folding kinetics of small proteins [47–49,51]. It has also been argued that topology and stability critically affect the folding rate [52–54]. The FDC mechanism relies essentially on the same ingredients although in a somewhat different form. Topology of the native state is linked to the distribution of the IS helices in sequence. This brings in the separation in sequence of the critical residues that is related to the contact order introduced in Ref. [55]. The signature of the key residues (spatial proximity in the native structure) is here supplanted by the minimal entropy criterion and the participation in the IS helices. Interestingly, from the DSSP files it turns out that the IS residues listed in Table I are in mutual contact. Therefore, since they meet the criterion used in Ref. [55] they contribute to the final value of the contact order.

As far as the second determinant of folding is concerned, in the FDC model stability is only partially accounted for by the folding and orientational probabilities β_i^n and γ_k^n. Other contributions, like the stabilities of the non-IS helices are neglected whereas long-range interactions among the microdomains are only in part and implicitly taken into consideration in the β_i^n and γ_k^n.

In this connection it must be noted that some approximations affect our results in that our picture relies on a simplified description of the interactions among colliding helices. In fact we are using the simplest version of the DC model where the microdomains undergo a process of free diffusion with suitable boundary conditions. More sophisticated pictures of the intramolecular collisions are conceivable in which the parameters β_i^n include the effect of activation barriers [18]. In a sense this implies that in the FDC model the role of long-range interactions is underestimated with respect to that of short-range interactions. Nonetheless the effects of this approximation are mitigated in the case that local forces give a predominant contribution to the definition of the folding pathway and to the formation of the majority of native contacts [56,57]. This seems to be the case for the all-α proteins examined in this work.

From the point of view of the intervening interactions, our data are also relevant to the debate about the relative weight of global and local forces in the folding process [56–59]. Detailed studies have stressed that the balance between these two kinds of interactions is strongly non uniform along protein sequences [60]. Likewise, our results suggest that local forces are predominant in the IS helices examined in this paper.

Recent works have stressed that another kind of balance, the one between native and non-native forces, determines the folding mechanism of helical proteins, and in particular its position within the continuum of models between the DC and the NC scenarios [61]. In the FDC model both balances are affected by the βs parameters that assign a variable relative weight to the native vs non-native interactions and to local vs global forces. This is a hint as to the fact that the FDC model may provide a unifying theoretical framework that shifts from the DC mode (in the case of very stable foldons) toward less hierarchical mechanisms with decreasing values of the β_i^ns. This issue will be investigated in a forthcoming paper. This feature of the FDC model is in full accord with the recent extended nucleus theory that similarly views the DC and NC schemes as manifestations of the same underlying mechanism [1,2].

The FDC model shows also that local biases for helical structures make it possible to preaverage fast degrees of freedom. More than that, the effectivity of the FDC model in reproducing the overall kinetics lends support to the minimal entropy criterion as a useful tool to achieve a substantial reduction of the relevant degrees of freedom and to focus on the critical residues by mere inspection of the protein sequence. In this respect, the FDC scenario represents a significant advancement toward the extension of the Anfinsen's thermodynamic hypothesis (sequence determines the native structure) to the realm of kinetics (sequence determines the rate and pathways of folding). Finally, the FDC model stresses the role of the minimally frustrated segments of the protein's sequence in the rate-limiting stages of folding. From the point of view of the fundamental theory, the FDC model provides effective means to associate the appropriate minimalist models with real helical proteins [3]. The resulting reduced model possesses a simplified (but not oversimplified) energy landscape combining the relevant smoothness and roughness [62] that allow a fairly accurate description of the whole folding kinetics.

ACKNOWLEDGMENTS

The authors gratefully acknowledge discussions with L. Ambrosone, K. Dill, P-G. de Gennes, J. King, G. Moro, and D.L. Weaver who helped clarify some critical aspects. The present work was partially funded by the Italian Ministry for Research with a PRIN 2002 grant to M.C. and E.C. E.C. was also supported by a FISR 2002 grant. This work was also partially supported by grants of the Ministero dell' Università e della Ricerca Scientifica e Tecnologica (MURST) for the projects "Hydrolases from Thermophiles: Structure, Function and Homologous and Heterologous Expression" and "Development and implementation of algorithms for predicting protein structure." Partial funding from projects on Molecular Genetics and on Functional Genomics, delivered to R.C. by the Italian Centro Nazionale delle Ricerche (CNR), and a PNR 2001-2003 (FIRB) project on Postgenomics, are also acknowledged.

[1] A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **97**, 1525 (2000).

[2] S. Gianni, N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W.N. White, M. L. DeMarco, V. Daggett, and A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **100**, 13 286 (2003).

[3] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995).

[4] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[5] E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997).

[6] D. Thirumalai and D. K. Klimov, Curr. Opin. Struct. Biol. **9**, 197 (1999).

[7] V. Muñoz, Curr. Opin. Struct. Biol. **11**, 212 (2001).

[8] P. S. Kim and R. L. Baldwin, Annu. Rev. Biochem. **51**, 459 (1982).

[9] R. L. Baldwin and G. D. Rose, Trends Biochem. Sci. **24**, 26 (1999); *ibid.* **24**, 77 (1999).

[10] B. Zimm and J. K. Bragg, J. Chem. Phys. **31**, 526 (1959).

[11] V. Muñoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Nature (London) **390**, 196 (1997).

[12] V. Muñoz, E. R. Henry, J. Hofrichter, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **95**, 5872 (1998).

[13] D. K. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **97**, 2544 (2000).

[14] S. F. Betz, P. A. Liebman, and W. F. DeGrado, Biochemistry **36**, 2450 (1997).

[15] C. E. Schafmeister and R. M. Stroud, Curr. Opin. Biotechnol. **9**, 350 (1998).

[16] Y. Fezoui, D. L. Weaver and J. J. Osterhout, Proc. Natl. Acad. Sci. U.S.A. **91**, 3675 (1994).

[17] M. Karplus and D. L. Weaver, Nature (London) **260**, 404 (1976).

[18] M. Karplus and D. L. Weaver, Biopolymers **18**, 1421 (1979).

[19] M. Karplus and D. L. Weaver, Protein Sci. **3**, 650 (1994).

[20] M. Compiani, P. Fariselli, P-L. Martelli, and R. Casadio, Proc. Natl. Acad. Sci. U.S.A. **95**, 9290 (1998).

[21] R. Casadio, M. Compiani, P. Fariselli, and P. L. Martelli, ISMB **7**, 66 (1999).

[22] M. Compiani, P. Fariselli, and R. Casadio, Phys. Rev. E **55**, 7334 (1997).

[23] P.-G. de Gennes, J. Stat. Phys. **12**, 463 (1975).

[24] D. J. Tobias, J. E. Mertz, and C. L. Brooks III, Biochemistry **30**, 6054 (1991).

[25] R. Gilmanshin, S. Williams, R. H. Callender, W. H. Woodruff, and R. B. Dyer, Proc. Natl. Acad. Sci. U.S.A. **94**, 3709 (1997).

[26] P. A. Thompson, W. A. Eaton, and J. Hofrichter, Biochemistry **36**, 9200 (1997).

[27] W. A. Eaton, V. Muñoz, P. A. Thompson, E. R. Henry, and J. Hofrichter, Acc. Chem. Res. **31**, 745 (1998).

[28] G. Hummer, A. E. Garcia, and S. Garde, Proteins: Struct., Funct., Genet. **42**, 77 (2001).

[29] V. Smirnov, *Cours de Mathematiques Superieures* (Mir Editions, Moscow, 1970) Vol. II.

[30] D. Bashford, D. L. Weaver, and M. Karplus, J. Biomol. Struct. Dyn. **1**, 1243 (1984).

[31] K. K. Yapa and D. L. Weaver, J. Phys. Chem. **100**, 2498 (1996).

[32] R. V. Pappu and D. L. Weaver, Protein Sci. **7**, 480 (1998).

[33] C. A. Rohl and R. L. Baldwin, Biochemistry **36**, 8435 (1997).

[34] V. Muñoz and L. Serrano, Nat. Struct. Biol. **1**, 399 (1994).

[35] M. A. Jimenez, M. Bruix, C. Gonzalez, F. J. Blanco, J. L. Nieto, J. Herranz, and M. Rico, Eur. J. Biochem. **211**, 569 (1993).

[36] M. T. Reymond, G. Merutka, H. J. Dyson, and P. E. Wright, Protein Sci. **6**, 706 (1997).

[37] H. Frauenfelder and P. G. Wolynes, Phys. Today **47**, 58 (1985).

[38] J. D. Hirst and C. L. Brooks III, Biochemistry **34**, 7614 (1995).

[39] P. A. Jennings and P. E. Wright, Science **262**, 892 (1993).

[40] H. Roder, G. A. Elöve, and S. W. Englander, Nature (London) **335**, 700 (1988).

[41] R. E. Burton, J. K. Myers, and T. G. Oas, Biochemistry **37**, 5337 (1998).

[42] P. Thompson, V. Muñoz, G. S. Jas, E. R. Henry, W. A. Eaton, and J. Hofrichter, J. Phys. Chem. B **104**, 378 (2000)).

[43] C. L. Brooks III, J. Phys. Chem. **100**, 2546 (1996).

[44] S. A. Islam, M. Karplus, and D. L. Weaver, J. Mol. Biol. **318**, 199 (2002).

[45] M-H. Yu and J. King, Proc. Natl. Acad. Sci. U.S.A. **81**, 6584 (1984).

[46] A. R. Panchenko, Z. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **93**, 2008 (1996).

[47] A. R. Viguera, V. Villegas, F. X. Aviles, and L. Serrano, Folding Des. **2**, 23 (1996).

[48] F. Chiti, N. Taddei, P. Webster, D. Hamada, T. Fiaschi, G. Ramponi, and C. M. Dobson, Nat. Struct. Biol. **6**, 380 (1999).

[49] N. Taddei, F. Chiti, T. Fiaschi, M. Bucciantini, C. Capanni, M. Stefani, L. Serrano, C. M. Dobson, and G. Ramponi, J. Mol. Biol. **300**, 633 (2000).

[50] S. E. Jackson, Folding Des. **3**, R81 (1998).

[51] J. K. Myers, and T. G. Oas, Nat. Struct. Biol. **8**, 552 (2001).

[52] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson, Nat. Struct. Biol. **6**, 1005 (1999).

[53] D. Baker, Nature (London) **405**, 39 (2000).

[54] A. R. Dinner and M. Karplus, Nat. Struct. Biol. **8**, 21 (2001).

[55] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).

[56] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).

[57] R. Unger and J. Moult, J. Mol. Biol. **259**, 988 (1996).

[58] J. G. Saven and P. G. Wolynes, J. Mol. Biol. **257**, 199 (1996).

[59] C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **34**, 282 (1999).

[60] V. J. Hilser, D. Dowdy, T. G. Oas, and E. Freire, Proc. Natl. Acad. Sci. U.S.A. **95**, 9903 (1998).

[61] Y. Zhou, and M. Karplus, Nature (London) **401**, 400 (1999).

[62] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).