

An entropy criterion to detect minimally frustrated intermediates in native proteins

MARIO COMPIANI*[†], PIERO FARISELLI[‡], PIER LUIGI MARTELLI[‡], AND RITA CASADIO[‡]

*[†]Dipartimento di Scienze Chimiche, Università di Camerino, Via S. Agostino 1, 62032 Camerino MC, Italy; and [‡]Laboratorio di Biofisica, Università di Bologna, Via Irnerio 42, 40126 Bologna, Italy

Edited by Peter G. Wolynes, University of Illinois at Urbana-Champaign, Urbana, IL, and approved May 29, 1998 (received for review October 2, 1997)

ABSTRACT The analysis of the information flow in a feed-forward neural network suggests that the output of the network can be used to compute a structural entropy for the sequence-to-secondary structure mapping. On this basis, we formulate a minimum entropy criterion for the identification of minimally frustrated traits with helical conformation that correspond to initiation sites of protein folding. The entropy of protein segments can be viewed as a nucleation propensity that is useful to characterize putative regions where folding is likely to be initiated with the formation of stretches of α -helices under the predominant influence of local interactions. Our procedure is successfully tested in the search for initiation sites of protein folding for which independent experimental and computational evidence exists. Our results lend support to the view that folding is a hierarchical event in which, in harmony with the minimal frustration principle, the final conformation preserves structural modules formed in the early stages of the process.

Quite recently, the advent of powerful experimental and computational approaches (1–13) have led to an upsurge of interest in the pathways of protein folding. Such pathways are usually investigated by direct observation of the forward process leading from the shapeless molecule to the final structural organization. Here, we take a backward approach to the problem and show that the study of native structures of globular proteins sheds light on the chronology of structure formation and ultimately provides a test of the principle of minimal frustration (14, 15). In this context, one of the basic questions that is being posed is whether dissecting the fold of proteins in a hierarchy of structures, besides being an instrumental way to handle with their architectural complexity, reflects some fundamental feature of protein folding. Furthermore, it has been argued that the folding process must be parallelized and must avoid nonproductive pathways that end up in low energy nonnative conformations (10, 12, 16, 17). Such dead ends may be the typical fate of protein folding because of the fact that proteins are frustrated systems, i.e., systems for which the simultaneous minimization of all interaction energies is impossible (15). Frustration entails that proteins are characterized by a rugged energy landscape in which a huge number of relative minima corresponds to misfolded structures. The most remarkable effect on the kinetics of folding is that the random search for the lowest energy structure takes inordinately large times that exceed by several orders of magnitude the folding times of real proteins (Levinthal's paradox) (18). The paradox can be circumvented by assuming that the inevitable frustration of proteins is reduced by the existence of some local structures, dictated

essentially by short range interactions, that are compatible with the overall fold of the protein in the native state (principle of minimal frustration; refs. 14 and 15). Stated in this form the requirement of minimal frustration is consistent with a model of protein folding according to which the early steps of folding occur, before the rate-limiting step that involves the folding nucleus (19, 20), in the initiation sites (10) under the predominant influence of local forces (10, 11, 13, 19, 21–34). Thus in the early stages of folding, local search involves residues close to each other in sequence and results in the formation of stretches of secondary structure that may possibly be stabilized by nonspecific long-range interactions in the form of hydrophobic effects or the formation of a molten globule (10, 12, 20, 23, 26, 35–37). Our interest is mainly focused on the formation of the earliest traits of helical secondary structure that are supposed to be conserved throughout the remainder of the folding process (11, 13, 19, 29, 30, 35, 38) and form the core of the nucleation sites stabilized later. In particular, we pursue the aim of testing the hypothesis that α -helices may function as independent “seeds for folding” (38).

The present work relies on the notion that local interactions are a necessary prerequisite for both the formation of native secondary structure and the foldability of proteins (13, 35, 39, 40). This feature justifies the application of a method that is primarily sensitive to local interactions to detect α -helix fragments that are viable candidates for the role of minimally frustrated initiation sites. To this end, we take advantage of recent progress in the prediction of protein secondary structures from the residue sequence with neural networks (41) to estimate the nucleation propensity of traits of the protein molecule. We use a feed-forward neural network with standard backpropagation learning algorithm. In the present set up, more thoroughly described in refs. 41 and 42, the current input pattern \mathcal{F} is generated by a sliding window, 13 residues in length, scanning the whole residue sequence. The output layer contains two real-valued output neurons, with activations $o_g^i \in [0, 1]$, $i = 1, 2$. Each of them codes for a specific structural class σ_i (α -helix, non- α -helix) to be ascribed to the residue lying in the central position of the current input \mathcal{F} (for brevity, the structure will be thought of as being assigned to the pattern \mathcal{F}). Between the input and the output layers there is an additional hidden layer comprising two neurons.

The essential premise to our method is that the network reconstructs a mapping from the space of amino acid sequences to the space of the secondary structures that is equivalent, in some average sense detailed below, to the real folding process and can be viewed as an implementation of Anfinsen's hypothesis (43). It can be stated that the neural network mimics protein folding in that the mapping has been proved to be equivalent to a model of the early stages of protein folding based on the kinetic equations of contact variables (44). From a statistical perspective, one can argue that the network is sensitive to statistical regularities (45), and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/959290-5\$2.00/0
PNAS is available online at www.pnas.org.

This paper was submitted directly (Track II) to the *Proceedings* office.
[†]To whom reprint requests should be addressed. e-mail: compiani@camserv.unimc.it.

its application in the search for initiation sites can be justified firstly, by analyzing the flow of information in the neural network and, secondly, by making a conjecture as to the general mechanism of folding. The starting point is a theorem that ensures the convergence of the outputs $o_{j_g}^i$ of backpropagation neural networks to the conditional probabilities $P(\sigma_i | \mathcal{F})$ that the input \mathcal{F} is found in the structure σ_i (46). The statistical meaning of the $o_{j_g}^i$ is to be contrasted with Anfinsen's thermodynamic hypothesis (43) that can be cast in the form

$$P(\sigma_n | \mathcal{F}, \mathcal{C}) = 1 \quad [1]$$

where \mathcal{C} indicates the remainder of the protein sequence, or context, and σ_n is the native structure. It is evident that the output of the perceptron captures less information than is needed for the unambiguous assignment of the local structure. As expected, the neglect of relevant pieces of information contained in \mathcal{C} introduces some degree of uncertainty. In more rigorous terms, the relationship

$$o_{j_g}^i = P(\sigma_i | \mathcal{F}) = \sum_{\mathcal{C}} P(\sigma_i | \mathcal{F}, \mathcal{C}) P(\mathcal{C}), \quad \forall \sigma_i \quad [2]$$

shows that the structural assignment, carried out by the network on the basis of the bare pattern \mathcal{F} , relies on an average over all contexts (more precisely, over the contexts contained in the training set). Even when the network is working at its best, the approximation entailed by the averaging procedure is the main performance limiting factor (45) for those patterns whose essential determinants of the native secondary structure reside in the tertiary structure of the protein (47, 48). Conversely, averaging has little effect on the patterns \mathcal{F}_r , henceforth termed reliable, whose secondary structure is largely independent of the context

\mathcal{C} . In fact, if we express the defining property of reliable patterns as

$$\partial P(\sigma_i | \mathcal{F}_r, \mathcal{C}) / \partial \mathcal{C} = 0, \quad \forall \sigma_i \quad [3]$$

and make use of Eq. 1, the above average over contexts simplifies to $\sum_{\mathcal{C}} P(\sigma_n | \mathcal{F}_r, \mathcal{C}) P(\mathcal{C}) = P(\sigma_n | \mathcal{F}_r) \sum_{\mathcal{C}} P(\mathcal{C}) = 1$, which ultimately implies $P(\sigma_n | \mathcal{F}_r) = 1$. However, under actual working conditions, it is to be expected that only approximate convergence is realized by the perceptron (46), i.e., $o_{j_g}^i \approx P(\sigma_i | \mathcal{F})$. In this case, the reliable patterns are more realistically defined by the condition

$$P(\sigma_n | \mathcal{F}_r) \approx o_{j_g}^i \approx 1, \quad [4]$$

which, in the following, will be referred to as the local version of the thermodynamic hypothesis. Eq. 4 provides a useful alternative definition of reliable patterns in terms of the output of the network.

A more practical way to implement the criterion of Eq. 4 is to define the entropy of the output vector

$$S_{j_g} = - \sum_{i=1}^2 P(\sigma_i | \mathcal{F}) \ln P(\sigma_i | \mathcal{F}) \approx - \sum_{i=1}^2 o_{j_g}^i \ln o_{j_g}^i, \quad [5]$$

and turn Eq. 4 into a minimal entropy criterion,

$$S_{j_g} \approx 0. \quad [6]$$

Some remarks about the physical content of Eq. 3 are in order to elucidate the meaning of the minimal entropy criterion of Eq. 6. Our selection of the reliable helices through Eq. 3 bears on the question about the relative weight of local and long-range (tertiary) contacts in the stabilization of the secondary structure. As

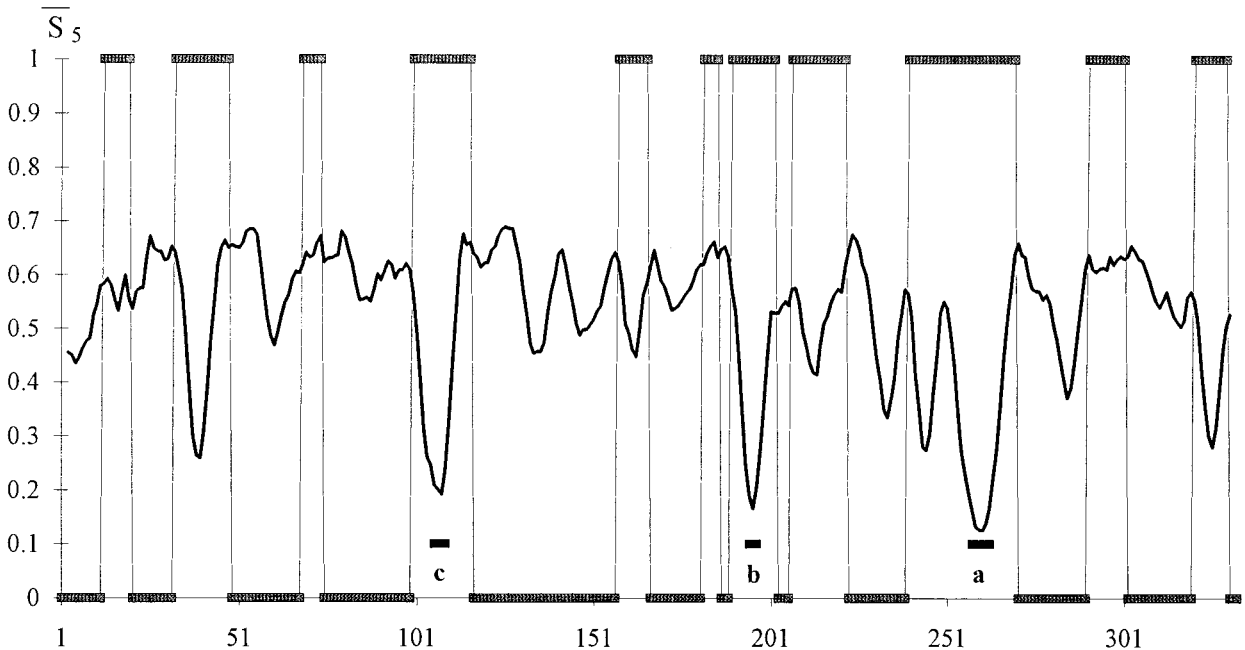


FIG. 1. Profile of the smoothed entropy \bar{S}_5 (continuous curve) for the protein 1GD1_O. The computation of \bar{S}_5 involves the prediction of the secondary structure of the pattern \mathcal{F} and, subsequently, the evaluation of S_{j_g} (Eq. 5) for each of the 334 residues of the protein. We have used a two-output network, designed to discriminate between α -helices and non- α structures. For crossvalidation of the network, we have used a training set comprising 300 nonhomologous globular proteins with <25% of sequence homology. Furthermore, we have normalized the actual output of the network to minimize the effect of small deviations of the $o_{j_g}^i$ from the ideal condition where they sum to unity. Next, a local average of S_{j_g} has been performed over a movable window five-residue long and centered on the current residue. The abscissa indicates the running position of the central residue along the protein backbone. We have superimposed a binary step function with nonzero plateaus spanning the regions that the network predicts in a α -helix structure. The zero plateaus of the step function (on the abscissa) indicate the segments that the neural network classifies in nonhelical structure. According to the minimal entropy criterion, the lowest minima of \bar{S}_5 , corresponding to helical conformations, hallmark the residues that are likely to belong in the initiation sites of folding of the protein. The boundaries of each initiation site have been fixed by admitting deviations in \bar{S}_5 of 0.05 around each local minimum. The three sites (c, b, a) are indicated by black bars under the corresponding entropy minimum and have been redrawn in Fig. 3 as boxes of appropriate length along the protein backbone.

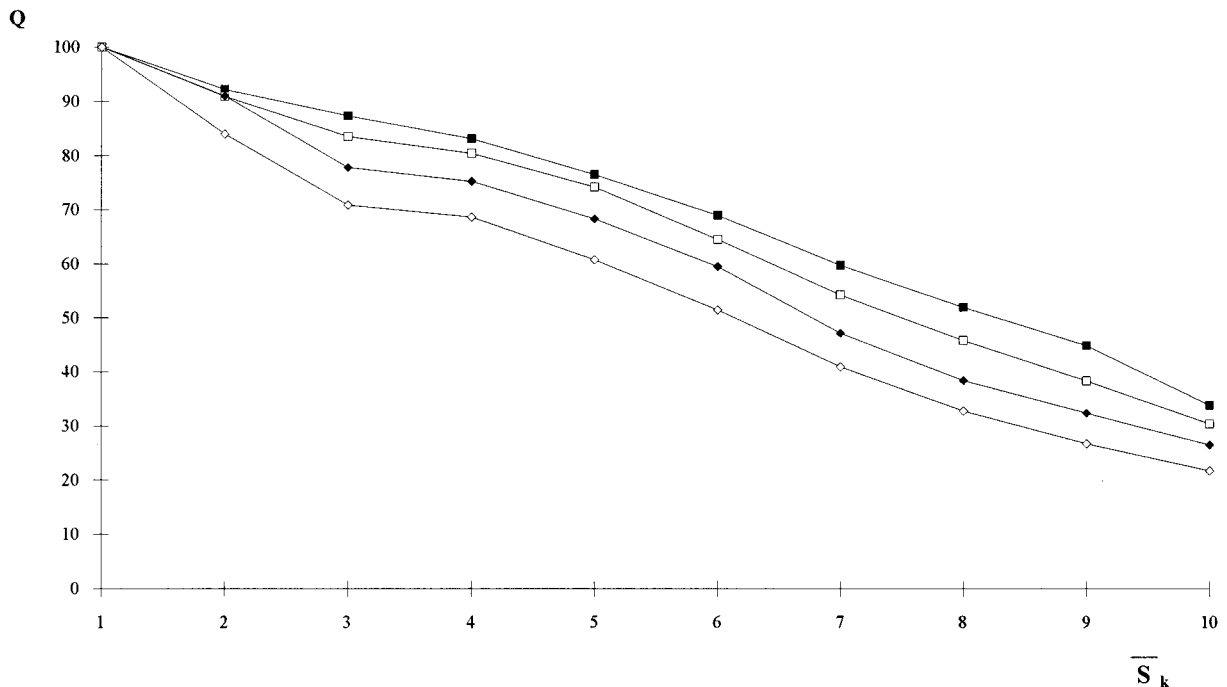


FIG. 2. Correlation between entropy measures and the performance Q of the network. The accuracy Q for each entropy interval is the percentage of the correctly predicted patterns normalized to the total number of patterns of the training set whose entropy lies in that same entropy interval. The entropy axis reports the average entropies \bar{S}_k , with $k = 1,3$ (open \square and filled \blacksquare square), $k = 5,7$ (open \diamond and filled \blacklozenge diamond). Note that S_g of Eq. 5 coincides with \bar{S}_1 . The range of entropy values has been discretized for convenience by dividing the allowed range $[0, \ln 2]$ into 10 equal intervals. The plot shows that there exists a strong correlation between the correctness of the network's prediction and low entropy segments up to five residues in length. Accordingly, we have used \bar{S}_5 in the search for the initiation sites reported in Fig. 3.

far as the reliable traits with helical conformation are concerned, Eq. 3 is compatible with two alternative views but the properties of the reliable patterns thus selected change correspondingly. In the first case, the reliable pattern is context-independent and intrinsically stable so that it can assume its native structure in solution even when it is excised from the protein. The existence of such segments is investigated in (11–13, 23, 28, 31, 33, 40, 49–58). Reliable patterns conforming to the second possibility are weakly context-dependent in that they are expected to be stabilized by long-range generic contributions (10, 12, 20, 23, 26, 28, 35–37) depending on compositional properties of the sequence rather than on specific tertiary interactions. Such properties supposedly are associated with those features shared by most of the sequences of the database (hence Eq. 3 is fulfilled) and correspond to the so-called self-averaging features in the terminology of spin glass theory (15).

Therefore, in agreement with the hierarchical models of folding (13, 22–26, 38), we argue that it is reasonable to equate the reliable patterns with the initiation sites where folding is started in the early stages, because they share the distinguishing feature, formalized in Eq. 3, that the attendant local structure is formed with the minor contribution of tertiary interactions. This conjecture enables us to convert the search of the initiation sites of folding to the problem of assessing, for any input \mathcal{F} , the conformity of the σ'_g with Eq. 4.

Accordingly, a useful test of the minimal entropy criterion can be carried out by comparing our predictions with data from the literature concerning the existence and location of folding intermediates and initiation sites. To select the initiation sites of folding, we draw plots of S_g along the residue sequence of each protein. After the signal is regularized, by averaging S_g over k neighboring residues, we search the entropy profile for the most pronounced minima corresponding to native helices that, supposedly, are correctly identified by the network. Fig. 1 exemplifies the trend of the average entropy \bar{S}_k with $k = 5$, that is the requisite span compatible with the shortest stretches of α -helices. Fig. 2 confirms that we may be confident that the

entropy-driven selection of the helices predicted by the neural network is a reliable procedure to identify native structures of the protein. The diagram makes clear that the lower the values of \bar{S}_k (up to segments of length $k = 5$) the more exact is the prediction of the corresponding patterns, indicating that the probability density $P(\sigma | \mathcal{F}_r)$, as approximated by the σ'_g , peaks on the correct (i.e., native) structure σ_n . This agreement is not surprising because our entropy provides much the same information as the reliability index discussed, for example, in ref. 59. As a crucial benchmark, we have selected the folding nuclei, determined by energy minimization procedures (11, 12, 49), that are expected to be intrinsically stable. We also have taken into consideration a set of proteins for which the participation of some tagged residues in intermediate stages of folding has been proved experimentally (3, 35, 60, 61, 63, 64) or supported by computer simulations (65). The data collected in Fig. 3 demonstrate that there is substantial agreement between our findings and the results obtained with the alternative methods. Notably, a glance at the entropy values listed in the caption of Fig. 3 shows that most of the selected helices of type *a* have entropies coinciding with or close to the absolute minimum of the \bar{S}_5 profile.

In conclusion, we note that our implementation of the minimal entropy criterion shares the simplicity of all neural approaches in that one needs a minimum amount of *a priori* information to achieve an effective extraction of statistical features. A more cumbersome analysis relying on an entropy concept can be found in ref. 66. The physical meaning attached to Eq. 3 justifies our contention that the lowest minima of the entropy plot are appropriate markers for minimally frustrated regions, where local cooperativity is the basic determinant of early helical structures that have inherent harmony with other structural features (15). Furthermore, the compatibility of the helical structure of our initiation sites with virtually any context of the training set makes the reliable patterns eligible as building blocks of secondary structure that are formed early and are retained throughout the later stages of folding (12, 23)

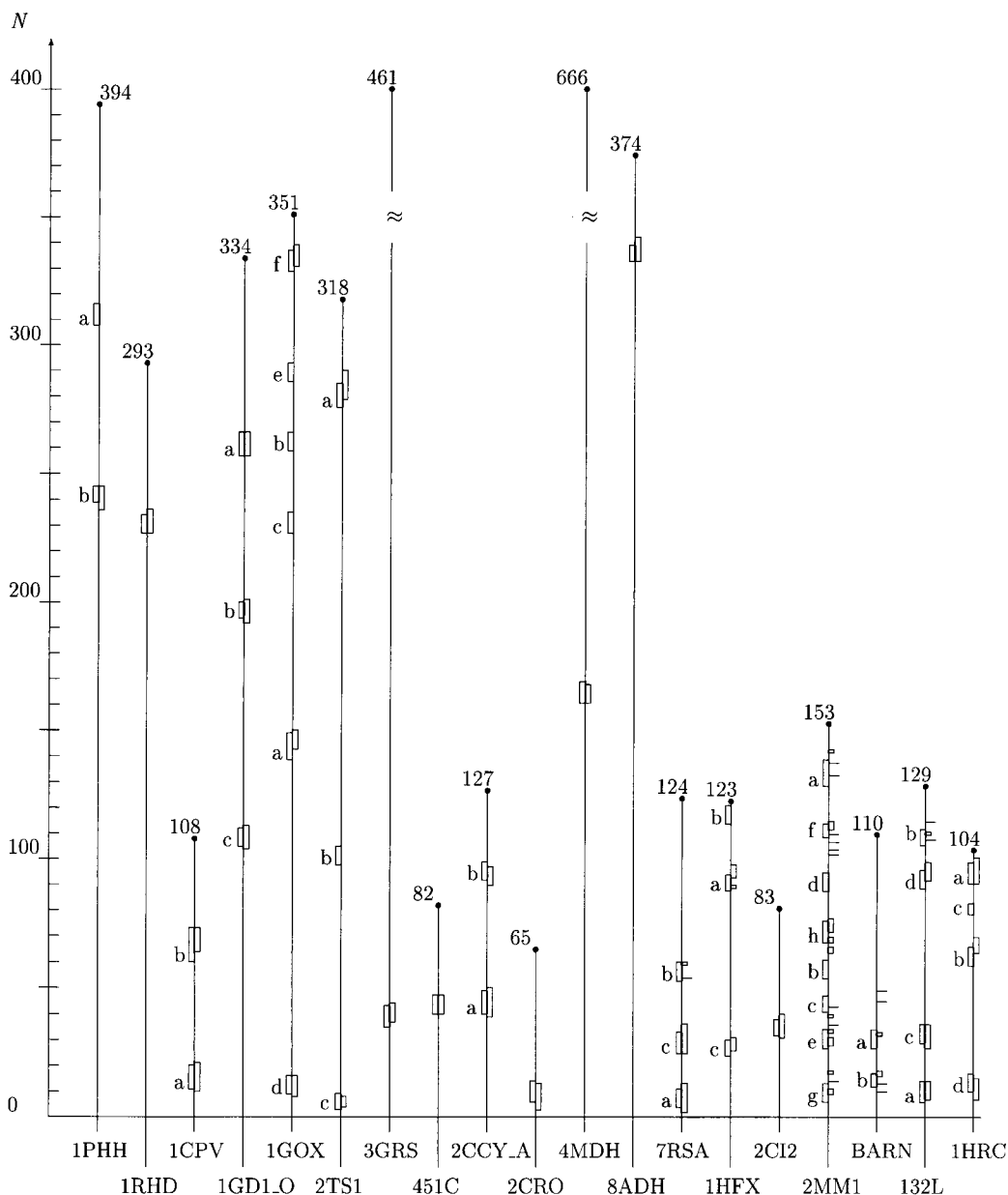


FIG. 3. Initiation sites of folding assuming helical structure determined by means of the minimal entropy criterion and other methods. On the abscissa, we have listed the Brookhaven codes of the proteins examined. The grid on the N -axis marks the position of the amino acids along the protein backbones, which have been skeletonized as vertical filaments of appropriate length. For convenience, the total number of residues of each protein is shown at the upper end of each filament. The boxes attached to each protein backbone highlight the location and extent of the earliest stretches of helical structure being formed during the folding process. The boxes on the left side of each filament have been positioned in the regions around the lowest minima of the average entropy \bar{S}_5 according to the criterion described in the caption of Fig. 1. The boxes or tick marks (in the case of single residues) on the right side of each backbone visualize the initiation sites as they result from experimental and computational works. Our segments are labeled with letters indicating ascending values of the entropy minima. The data for the first 12 proteins (1PHH, 1RHD, 1CPV, 1GDI_O, 1GOX, 2T51, 3GRS, 451C, 2CCY_A, 2CRO, 4MDH, and 8ADH) are drawn from ref. 49; the experimental results for 7RSA have been drawn from ref. 11 for site a and c and ref. 62 for site b . The site a also has been predicted in ref. 12 in the same position as determined in ref. 11. Helix c (but the last two residues) also has been shown to be present in the early folding intermediate identified in ref. 62. The experimentally detected sites for 1HFX, 2CI2, 2MM1, BARN (barnase), 132L, and 1HRC are drawn, respectively, from refs. 35, 60, 3, 61, 34, and 64. Our initiation site for 2CI2 includes the first two residues that comprise the folding nucleus according to ref. 65. The experimental data reported in ref. 3 for 2MM1 and ref. 34 for 132L essentially confirm the previous analysis in refs. 13 and 63, respectively. In the following, we list the range of variation δS of the entropy in the entropy profile and the site entropy S for each protein examined. 1PHH: $\delta S = 0.224-0.690$, $S(a) = 0.224$, $S(b) = 0.253$; 1RHD: $\delta S = 0.173-0.687$, $S = 0.173$; 1CPV: $\delta S = 0.237-0.677$, $S(a) = 0.237$, $S(b) = 0.305$; 1GDI_O: $\delta S = 0.125-0.689$, $S(a) = 0.125$, $S(b) = 0.166$, $S(c) = 0.192$; 1GOX: $\delta S = 0.288-0.691$, $S(a) = 0.288$, $S(b) = 0.290$, $S(c) = 0.294$, $S(d) = 0.307$, $S(e) = 0.311$, $S(f) = 0.315$; 2T51: $\delta S = 0.227-0.687$, $S(a) = 0.256$, $S(b) = 0.273$, $S(c) = 0.284$; 3GRS: $\delta S = 0.154-0.688$, $S = 0.245$; 451C: $\delta S = 0.142-0.688$, $S = 0.237$; 2CCY_A: $\delta S = 0.094-0.672$, $S(a) = 0.094$, $S(b) = 0.178$; 2CRO: $\delta S = 0.259-0.637$, $S = 0.259$; 4MDH: $\delta S = 0.166-0.690$, $S = 0.261$; 8ADH: $\delta S = 0.153-0.690$, $S = 0.292$; 7RSA: $\delta S = 0.326-0.689$, $S(a) = 0.350$, $S(b) = 0.508$, $S(c) = 0.533$; 1HFX: $\delta S = 0.341-0.691$, $S(a) = 0.435$, $S(b) = 0.490$, $S(c) = 0.520$; 2CI2: $\delta S = 0.300-0.684$, $S = 0.428$; 2MM1: $\delta S = 0.264-0.686$, $S(a) = 0.264$, $S(b) = 0.378$, $S(c) = 0.451$, $S(d) = 0.460$, $S(e) = 0.466$, $S(f) = 0.524$, $S(g) = 0.616$, $S(h) = 0.616$; BARN: $\delta S = 0.261-0.689$, $S(a) = 0.376$, $S(b) = 0.558$; 132L: $\delta S = 0.190-0.688$, $S(a) = 0.301$, $S(b) = 0.494$, $S(c) = 0.556$, $S(d) = 0.647$; 1HRC: $\delta S = 0.244-0.688$, $S(a) = 0.296$, $S(b) = 0.374$, $S(c) = 0.586$, $S(d) = 0.644$. The helices marked a correspond to helical structures with the lowest entropy. They coincide with or are close to the absolute minimum of the entropy profile (see text). We have included as many segments as needed to scan all the α -helices assigned to each protein in the literature. Our numeration for 2CI2 is shifted ahead by 19 residues with respect to the one in ref. 60.

irrespective of the subsequent adjustments of the tertiary structure.

The remarkable agreement of our data with the results in (49) indicates that self-stabilizing segments are successfully captured by our criterion. Notably, experimental investigations have shown that some of the initiation sites predicted in Fig. 3 settle autonomously in the helical conformation even when they belong to isolated peptides in solution. This is the case of the helices *a*, *b*, and *c* of 7RSA studied in refs. 51–53 and the *a* helices of 2MM1 and 1HRC tested in refs. 54 and 55. However the 2-fold interpretation of our selection rule for reliable patterns (Eq. 3) does not allow one to infer that isolated reliable patterns are necessarily stable in solution. Examples of such segments for which the local helical propensity must be assisted by a suitable hydrophobic environment are the B and C helices of α -lactalbumin [entirely unfolded in water (35)] that correspond to the *d* and *b* traits of 1HFX in Fig. 3.

By contrast, tertiary interactions dictate more stringent and specific constraints to the conformation of unreliable patterns that are characterized by higher values of the entropy than the reliable ones. Turns appear to be good candidates for unreliable patterns in α -helical proteins on account of their role of floppy regions in which local effects are overcome by long-range interactions (67). Notably the existence of a mix of unreliable and reliable patterns has been shown to favor the foldability of proteins (40).

The authors wish to thank Dr. P. G. Wolynes and Dr. B. Rost for helpful comments on preliminary versions of the paper. Financial support to this work was provided by Ministero della Ricerca Scientifica e Tecnologica and a Target Project in Biotechnology from the Centro Nazionale delle Ricerche (CNR).

- Miranker, A., Robinson, C. V., Radford, S. E., Aplin, R. T. & Dobson, C. M. (1993) *Science* **262**, 896–900.
- Arcus, V. L., Vuilleumier, S., Freund, S. M. V., Bycroft, M. & Fersht, A. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 9412–9416.
- Jennings, P. A. & Wright, P. E. (1993) *Science* **262**, 892–896.
- Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1996) *J. Mol. Biol.* **257**, 430–440.
- Service, R. F. (1996) *Science* **273**, 29–30.
- Dandekar, T. & Argos, P. (1994) *J. Mol. Biol.* **236**, 844–861.
- Sali, A., Shakhnovich, E. & Karplus, M. (1994) *Nature (London)* **369**, 248–251.
- Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995) *Protein Sci.* **4**, 1167–1177.
- Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995) *J. Mol. Biol.* **252**, 460–471.
- Wright, P. E., Dyson, H. J. & Lerner, R. A. (1988) *Biochemistry* **27**, 7167–7175.
- Moult, J. & Unger, R. (1991) *Biochemistry* **30**, 3816–3824.
- Avbelj, F. & Moult, J. (1995) *Proteins Struct. Funct. Genet.* **23**, 129–141.
- Hughson, F. M., Wright, P. E. & Baldwin, R. L. (1990) *Science* **249**, 1544–1548.
- Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.
- Frauenfelder, H. & Wolynes, P. G. (1994) *Phys. Today* **47**, 58–64.
- Panchenko, A. R., Luthey-Schulten, Z. & Wolynes, P. G. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 2008–2013.
- Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 9029–9033.
- Levinthal, C. (1969) *Mössbauer Spectroscopy in Biological Systems*, eds. Debrunner, P., Tsiibris, J.-C. & Münck, E. (University of Illinois Press, Urbana, IL), pp. 22–24.
- Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *Biochemistry* **33**, 10026–10036.
- Christensen, H. & Pain, R. H. (1991) *Eur. Biophys. J.* **19**, 221–229.
- Fersht, A. R. (1993) *FEBS Lett.* **325**, 5–16.
- Kim, P. S. & Baldwin, R. L. (1990) *Annu. Rev. Biochem.* **59**, 631–660.
- Lesk, A. M. & Rose, G. D. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4304–4308.
- Go, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.
- Karplus, M. & Weaver, D. L. (1994) *Protein Sci.* **3**, 650–668.
- Ptitsyn, O. B. & Uversky, Y. N. (1994) *FEBS Lett.* **341**, 15–18.
- Udgaonkar, J. B. & Baldwin, R. L. (1988) *Nature (London)* **335**, 694–699.
- Saven, J. G. & Wolynes, P. G. (1996) *J. Mol. Biol.* **257**, 199–216.
- Weissman, J. S. & Kim, P. S. (1991) *Science* **253**, 1386–1392.
- Oas, T. G. & Kim, P. S. (1988) *Nature (London)* **336**, 42–48.
- Bai, Y., Sosnick, T. R., Mayne, L. & Englander, S. W. (1995) *Science* **269**, 192–197.
- Briggs, M. S. & Roder, H. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2017–2021.
- Roder, H., Elöve, G. A. & Englander, S. W. (1988) *Nature (London)* **335**, 700–704.
- Radford, S. E., Dobson, C. M. & Evans, P. A. (1992) *Nature (London)* **358**, 302–307.
- Chyan, C.-L., Wormald, C., Dobson, C. M., Evans, P. A. & Baum, J. (1993) *Biochemistry* **32**, 5681–5691.
- Chan, H. S. & Dill, K. A. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 6388–6392.
- Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 1942–1946.
- Presta, L. G. & Rose, G. D. (1988) *Science* **240**, 1632–1641.
- Socci, N. D., Bialek, W. S. & Onuchic, J. N. (1994) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **49**, 3440–3443.
- Unger, R. & Moult, J. (1996) *J. Mol. Biol.* **259**, 988–994.
- Fariselli, P., Compiani, M. & Casadio, R. (1993) *Eur. Biophys. J.* **22**, 41–51.
- Casadio, R., Fariselli, P., Taroni, C. & Compiani, M. (1996) *Eur. Biophys. J.* **24**, 165–178.
- Anfinsen, C. B. (1973) *Science* **181**, 223–230.
- Bohr, G. & Wolynes, P. G. (1992) *Phys. Rev. A At. Mol. Opt. Phys.* **46**, 5242–5247.
- Compiani, M., Fariselli, P. & Casadio, R. (1997) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **55**, 7334–7343.
- Bishop, C. M. (1994) *Rev. Sci. Instrum.* **65**, 1803–1832.
- Zhong, L. & Johnson, W. C., Jr. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4462–4465.
- Minor, D. L. & Kim, P. S. (1994) *Nature (London)* **371**, 264–267.
- Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1992) *Biochemistry* **31**, 10226–10238.
- Kim, P. S. & Baldwin, R. L. (1984) *Nature (London)* **307**, 329–334.
- Shoemaker, K. R., Fairman, R., Kim, P. S., York, E. J., Stewart, J. M. & Baldwin, R. L. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 391–398.
- Jimenez, M. A., Rico, M., Herranz, J., Santoro, J. & Nieto, J. L. (1988) *Eur. J. Biochem.* **175**, 101–109.
- Jimenez, M. A., Nieto, J. L., Herranz, J., Rico, M. & Santoro, J. (1987) *FEBS Lett.* **221**, 320–324.
- Waltho, J. P., Feher, V. A., Merutka, G., Dyson, H. J. & Wright, P. E. (1993) *Biochemistry* **32**, 6337–6347.
- Kuroda, Y. (1993) *Biochemistry* **32**, 1219–1224.
- Muñoz, V. & Serrano, L. (1994) *Nat. Struct. Biol.* **1**, 399–409.
- Baldwin, R. L. (1995) *Biophys. Chem.* **55**, 127–135.
- Bagby, S., Go, S., Inouye, S., Ikura, M. & Chakrabarty, A. (1998) *J. Mol. Biol.* **276**, 669–681.
- Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232**, 584–599.
- Fersht, A. R. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 10869–10873.
- Bycroft, M., Matouschek, A., Kellis, J. T., Jr., Serrano, L. & Fersht, A. (1990) *Nature (London)* **346**, 488–490.
- Udgaonkar, J. B. & Baldwin, R. L. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 8197–8201.
- Miranker, A., Radford, S. E., Karplus, M. & Dobson, C. M. (1991) *Nature (London)* **349**, 633–636.
- Jeng, M.-F., Englander, S. W., Elöve, G. A., Wand, A. J. & Roder, H. (1990) *Biochemistry* **29**, 10433–10437.
- Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996) *Nature (London)* **379**, 96–98.
- Kang, H. S., Kurochkina, N. A. & Lee, B. (1993) *J. Mol. Biol.* **229**, 448–460.
- Brunet, A. P., Huang, E. S., Huffiner, M. E., Loeb, J. E., Weltman, R. J. & Hecht, M. H. (1993) *Nature (London)* **364**, 355–358.