Theor Chem Acc (1999) 101:21–26
DOI 10.1007/s002149800m89

Theoretical
Chemistry Accounts

Regular article

Neural networks to study invariant features of protein folding*

M. Compiani[1], **P. Fariselli**[2], **P. L. Martelli**[2], **R. Casadio**[2]

[1] Dipartimento di Scienze Chimiche, Universitá di Camerino, Via S. Agostino 1, I-62032 Camerino MC, Italy
[2] Laboratorio di Biofisica, Universitá di Bologna, Via Irnerio 42, I-40126 Bologna, Italy

Received: 24 April 1998 / Accepted: 4 August 1998 / Published online: 11 November 1998

Abstract. Protein secondary structures result both from short-range and long-range interactions. Here neural networks are used to implement a procedure to detect regions of the protein backbone where local interactions have an overwhelming effect in determining the formation of stretches in α-helical conformation. Within the framework of a modular view of protein folding we have argued that these structures correspond to the initiation sites of folding. The hypothesis to be tested in this paper is that sequence identity beside ensuring similarity of the three-dimensional conformation also entails similar folding mechanisms. In particular, we compare the location and sequence variability of the initiation sites extracted from a set of proteins homologous to horse heart cytochrome *c*. We present evidence that the initiation sites conserve their position in the aligned sequences and exhibit a more reduced variability in the residue composition than the rest of the protein.

Key words: Initiation sites of protein folding – Neural networks – Self-stabilising helices – Homologous proteins – Positional invariance of initiation sites

1 Introduction

It is commonly accepted that protein folding poses a problem, the solution of which will lead to significant progress in our understanding of the relationship between structure and function in biomolecules. In the last decade the inherent difficulties in the simulation and theoretical analysis of folding have been circumvented, since it has been recognised that the native structure of proteins is the unique attractor of a dynamical process (Anfinsen's thermodynamic hypothesis) [1] and that the relevant initial conditions are specified in the residue sequence. As a consequence, a substantial impetus has been given to devising simplified substitutes of the underlying dynamics in terms of mappings from the space of amino acid sequences to the space of protein structures [2–5]. Concomitantly, thorough studies of the variety of protein structures have led to a standard classification of repetitive structural motifs and their rationalisation into a hierarchy of configurations [6].

In this context a natural question is whether one can adopt a bottom-up approach to the problem of protein folding to deduce relevant features of the dynamical process by inspection of the resulting native molecular structure [7]. Such a scenario is consistent with the picture of folding as a modular process that is started, in the so-called initiation sites, either by the early stabilisation of elements of secondary structure, prior to the appearance of tertiary contacts [8–12] or by the formation of transition states involving long-range interactions among the distant residues participating in the nucleus of folding [13]. Note that here the terms "short" and "long" represent the distance along the chain and not spatial separation. These early structures undergo subsequent accretion and are conserved within the native structure due to their intrinsic compatibility with the overall fold being stabilised later [11]. In general compatibility of partially formed substructures is a fundamental requirement to ensure the foldability of proteins. It has been suggested that this could be done by positing the existence of small modules independently capable of folding [8, 10, 11, 14]. The search for such structural modules has been grounded on geometric criteria [9] and more recently on statistical analysis of the energy landscape [15, 16]. The initiation sites referred to in this paper are stretches of α-helices that are formed in the early stages of folding and are conserved throughout the whole process [7]. Related notions that appeared in the recent literature are the folding nuclei described in Ref. [13], the "extended intermediates" and the building-blocks studied in Refs. [17–19], the foldons studied in Refs. [14, 16] and the notion of independent "seeds for folding" [20].

*Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: M. Compiani
(e-mail: compiani@camserv.unicam.it)

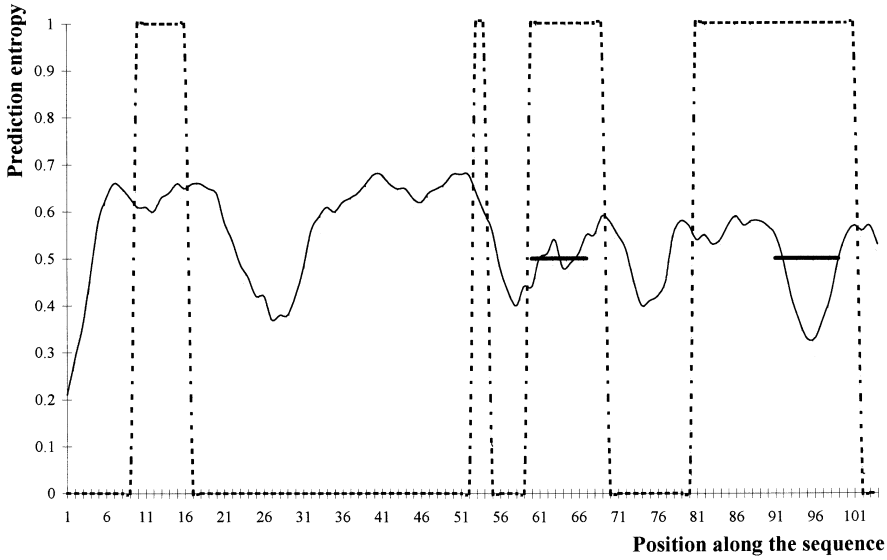

Fig. 1. Profile of the smoothed entropy $S_{\text{pred}}(l)$ (*continuous curve*) for horse heart cytochrome *c* (1HRC). Drawing the plot requires preliminary prediction of the secondary structure of each residue in the generic position $l(l = 1\text{–}104)$ and, subsequently, the evaluation of $S_{\text{pred}}(l)$ using Eq. (1). We have superimposed a binary step function (*dashed line*) with non-zero values centred on the regions that the network predicts in the α-helix structure. According to the minimal entropy criterion, the below-average minima of $S_{\text{pred}}(l)$ corresponding to helical conformations hallmark the residues that are likely to belong in the initiation sites of protein folding. The locations of the two initiation sites of the protein have been marked by *thick horizontal bars* around the appropriate entropy minimum. Note that they correspond to the two helices with lowest $S_{\text{pred}}(l)$

2 Searching for the initiation sites of folding in homologous proteins

To cope with the task of identifying the initiation sites of protein folding we take advantage of recent progress in the prediction of protein secondary structures from the bare residue sequence with neural networks [4, 5, 21]. Capitalizing on the statistical information captured by the network [21], we have introduced a structural entropy to measure the degree of independence of helical structures on the tertiary interactions [7]. This has led us to formulate in our previous work [7] a minimal entropy criterion for the identification of self-stabilising segments that are likely to serve as initiation sites and time-invariant building-blocks for individual proteins during the folding process. This procedure provides the basis for the present investigation, the main aim of which is to ask the question whether homologous protein share similar elementary modules of folding. The homologous proteins chosen to test our working hypothesis belong to the family of horse heart cytochrome *c* (PDB protein code: 1HRC). A total of 104 proteins (henceforth the testing set) with sequence identity larger than 45% have been selected from the data bank. This strict homology criterion ensures that the said proteins share the same three-dimensional structure [22, 23]. Sequence identity is calculated with reference to the aligned segments in the corresponding HSSP file ftp://ftp.EMBL-Heidelberg.de/pub/databases/hssp/1hrc.hssp [23]. The multiple sequence alignment has been carried out with the program CLUSTAL W [24]. We use the same feed-forward neural network as in Ref. [7] with a standard backpropagation learning algorithm. The only difference to mention is that the network has been trained on an extended set of 532 non-homologous proteins (homology $\leq 25\%$). The training set contained only one protein of the cytochrome family (1CCR) in common with the testing set. In the present set-up, which has been more thoroughly described in previous papers [4, 5, 7], the current input pattern is generated by a sliding window, 13 residues in length, scanning the N residues of the whole sequence. Each pattern is shifted one residue ahead with respect to the preceding one. The output layer contains two real valued output neurons, with activations $o^i(l) \in [0, 1]$, $i = 1, 2$ and $l = 1 - N$. Each of them codes for a specific structural class ($1 = \alpha$-helix, $2 =$ non α-helix) to be ascribed to the lth residue of the protein lying in the central position of the input pattern. In-between the input and output layers there is an additional intermediate layer with two real-valued neurons. The reference secondary structures are generated from the atomic crystallographic coordinates by means of the DSSP assignment [25].

The basic ingredient of our procedure is a measure of the local propensity for the helical structure. Here we use the results of Refs. [7, 21] and pursue this task by means of the entropy associated with the network's output [26] (henceforth prediction entropy)

$$S_{\text{pred}}(l) = -\sum_{i=1}^{2} o^i(l) \ln o^i(l) \tag{1}$$

The average entropy $\langle S_{\text{pred}} \rangle$ is defined as the average of $S_{\text{pred}}(l)$ over the entire sequence. Next, a local average of $S_{\text{pred}}(l)$ has been performed over a moveable window of five contiguous residues and centred on the current lth residue to generate the smoothed plot shown in Fig. 1. Hereafter, to simplify formalism, the smoothed prediction entropy will be simply referred to as predic-

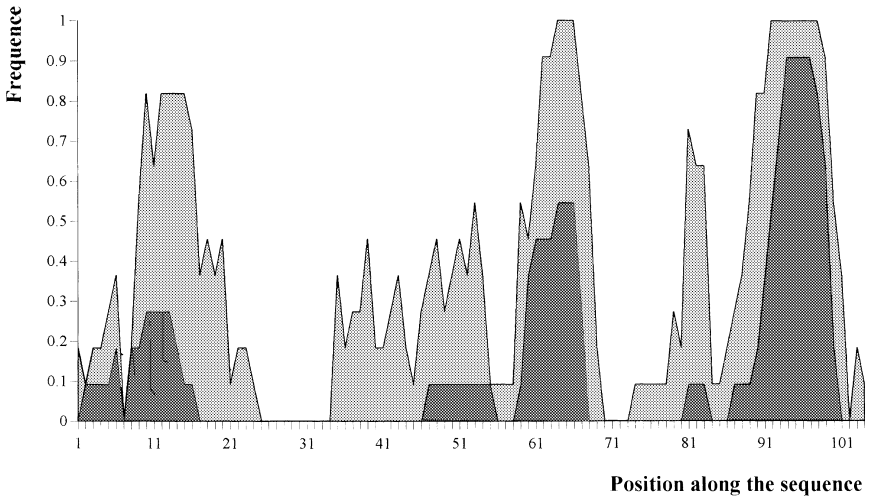

Fig. 2. Histograms of the α-helices predicted with the neural network in 11 homologous proteins (from the set $\mathscr{P}_{50}$) with sequence identity in the range 45–50%. The *light grey area* indicates the frequency of occurrence of α-helical structures for each position along the sequence. Similarly, the *dark grey area* refers to the frequency of occurrence of α-helices belonging to the initiation sites. Helices are quite well spread over the sequence whereas nucleation helices are mostly concentrated in the peaks around positions 64 and 94. In this figure, as in all subsequent ones, only aligned traits (without insertions) are displayed to facilitate the comparison between the data from sets $\mathscr{P}_{50}$ and $\mathscr{P}_{100}$

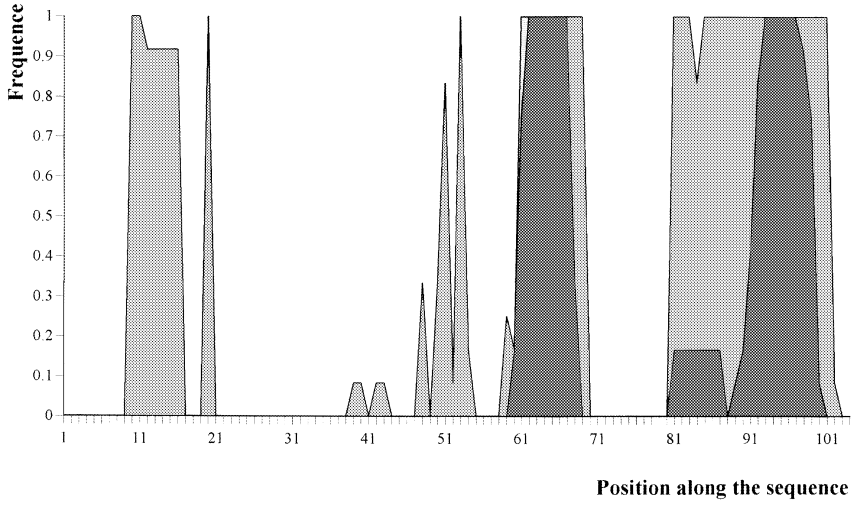

Fig. 3. Histograms of the α-helices predicted with the neural network in 12 homologous proteins (from the set $\mathscr{P}_{100}$) with sequence identity in the range 90–100%. The *light grey* and *dark grey areas* indicate the same as in Fig. 2. The concentration of nucleation helices in the two peaks already visible in Fig. 2 is more evident here

tion entropy and indicated by the symbol $S_{\text{pred}}(l)$. According to the minimal entropy criterion [7] we can determine the location of the initiation sites by looking at the residues with α-helical structure that fulfil the condition $S_{\text{pred}}(l) \leq \langle S_{\text{pred}} \rangle$ (see Fig. 1). For the purpose of the present investigation we grouped the residues from proteins with high and low sequence identity. Starting from the 104 proteins of the testing set two subsets were thus created, $\mathscr{P}_{100}$ including the residues from 12 proteins with 90–100% sequence identity and $\mathscr{P}_{50}$ of residues from 11 proteins with 45–50% sequence identity. The computation of $S_{\text{pred}}(l)$ has been carried out for each protein used to build $\mathscr{P}_{100}$ and $\mathscr{P}_{50}$ and, eventually, initiation sites have been individuated in each entropy plot (see Fig. 1) following the minimal entropy criterion. The next step consists in comparing the distribution along the sequences of the nucleation helices with the distribution of α-helices that do not belong to initiation sites. To carry out this analysis we have further split the subsets $\mathscr{P}_{100}$ and $\mathscr{P}_{50}$ as follows. By way of example, we have taken $\mathscr{P}_{100} = \mathscr{H}^{I}_{100} \cup \mathscr{H}^{nI}_{100} \cup \mathscr{C}_{100}$, where the subsets $\mathscr{H}^{I}_{100}$ and $\mathscr{H}^{nI}_{100}$ contain, respectively, residues belonging to helices that are or are not initiation sites. The subset $\mathscr{C}_{100}$ is comprised of all the residues in non-helical structures. A similar partition has been performed on the set $\mathscr{P}_{50}$. Note that the protein 1CCR of the training set is contained neither in $\mathscr{P}_{100}$ nor in $\mathscr{P}_{50}$. The results are shown in Figs. 2 and 3. It is quite apparent that the initiation sites are affected by a moderate spread and, furthermore, their location is remarkably conserved along the sequences of the proteins even when there is a non-negligible difference in sequence identity.

3 Conserved patterns in the initiation sites of folding

The second issue concerns the conservation of residues along the sequence. The basic idea is that if early intermediates play a role in driving the folding process and determining the native structure, they are expected to be screened better than average from indiscriminate variability. A related investigation is to be found in the work of Shakhnovich et al. [27]. To test this hypothesis we use a sequence entropy $S_{\text{seq}}(l)$ defined, on the aligned traits (without insertions) of the 104 homolo-

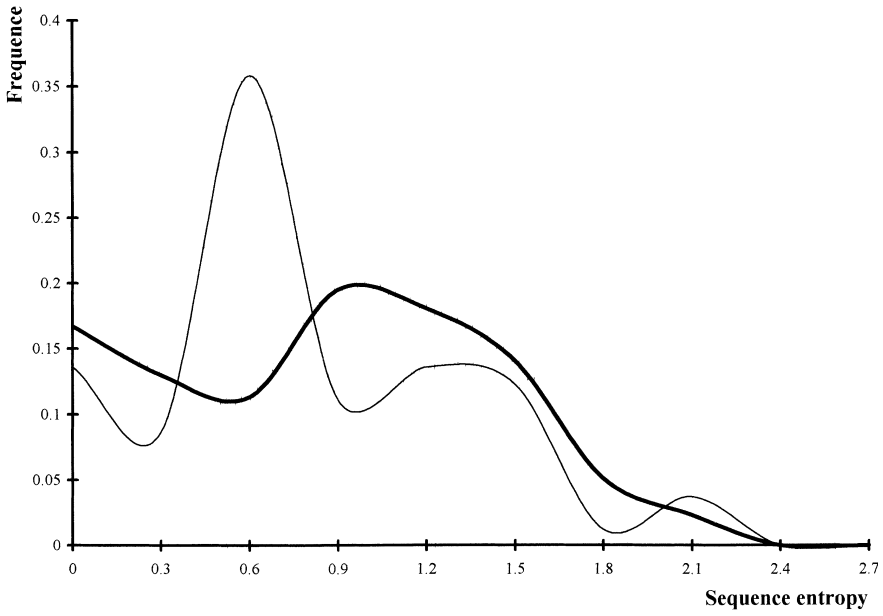

Fig. 4. Smoothed histograms of the sequence entropy $S_{\text{seq}}(l)$ of the 11 proteins used to build the set $\mathscr{P}_{50}$. The *thin line* displays the frequency of helices belonging to the initiation sites. The *thick line* refers to all the other helices not belonging to the initiation sites

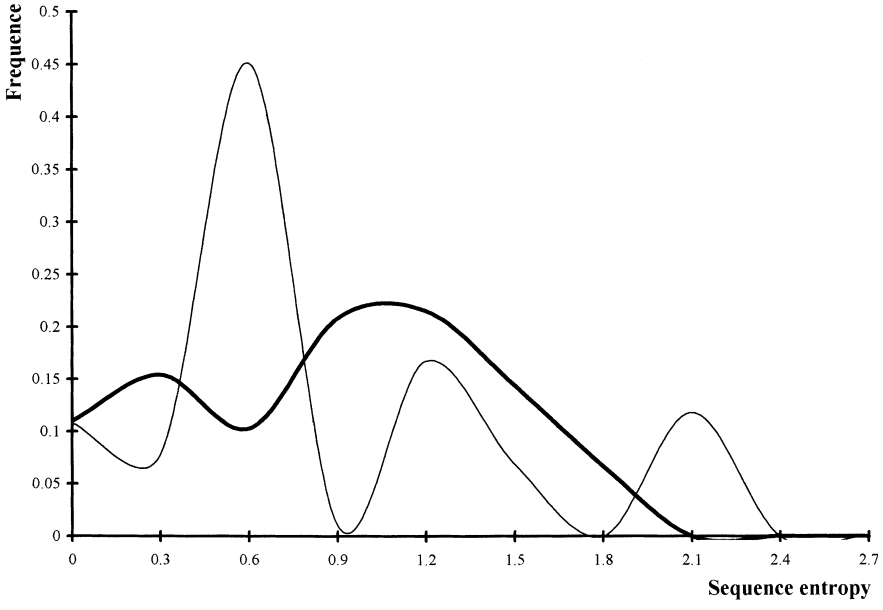

Fig. 5. Smoothed histograms of the sequence entropy $S_{\text{seq}}(l)$ of the 12 proteins used to build the set $\mathscr{P}_{100}$. The *thin* and *thick lines* indicate the same as in Fig. 4

gous proteins, by an equation similar to Eq. (1), where $o^i(l)$ is the probability of occurrence of the ith residue $(i = 1 - 20)$ in the lth position. Clearly conserved residues are characterized by low values of $S_{\text{seq}}(l)$. For the purpose of highlighting a possible relationship between conservatism and participation in the initiation sites of folding it is convenient to compare histograms for the entropy $S_{\text{seq}}(l)$ of the subsets $\mathscr{H}^I_{100}$, $\mathscr{H}^{nI}_{100}$, $\mathscr{H}^I_{50}$ and $\mathscr{H}^{nI}_{50}$. From Figs. 4 and 5 one can see that the initiation helices, compared with the remainder of the proteins, exhibit a less pronounced spread in $S_{\text{seq}}(l)$ and that their maxima are shifted towards smaller values of sequence entropy. This is a hint that the sequence of the nucleation helices is on average better conserved than in all the other regions of the protein molecule or, equivalently, that low values of $S_{\text{pred}}(l)$ tend to correlate with low values of $S_{\text{seq}}(l)$. In order to visualise this trend better we have drawn histograms showing the joint probability $P(S_{\text{seq}}, S_{\text{pred}})$. An example is given in Fig. 6 for the data relating to the set of nucleation helices $\mathscr{H}^I_{100}$. To obtain more quantitative information about the possible correlation between the two entropies we have calculated the correlation coefficient ρ [28] between the values of $S_{\text{seq}}(l)$ and $S_{\text{pred}}(l)$ for the residues of sets $\mathscr{H}^i_k$ and $\mathscr{C}_k$ $(i = nI, I$ and $k = 50, 100)$. The results are summarised in Table 1. The gap between the values of ρ associated with $\mathscr{H}^I_k$ and $\mathscr{H}^{nI}_k$ $(k = 50, 100)$ indicates that a linear relationship exists between the entropies $S_{\text{seq}}(l)$ and $S_{\text{pred}}(l)$ of the residues located in the initiation helices (the significance levels for $\rho = 0.410$ and $\rho = 0.165$ are, respectively, 0.0005 and 0.14).

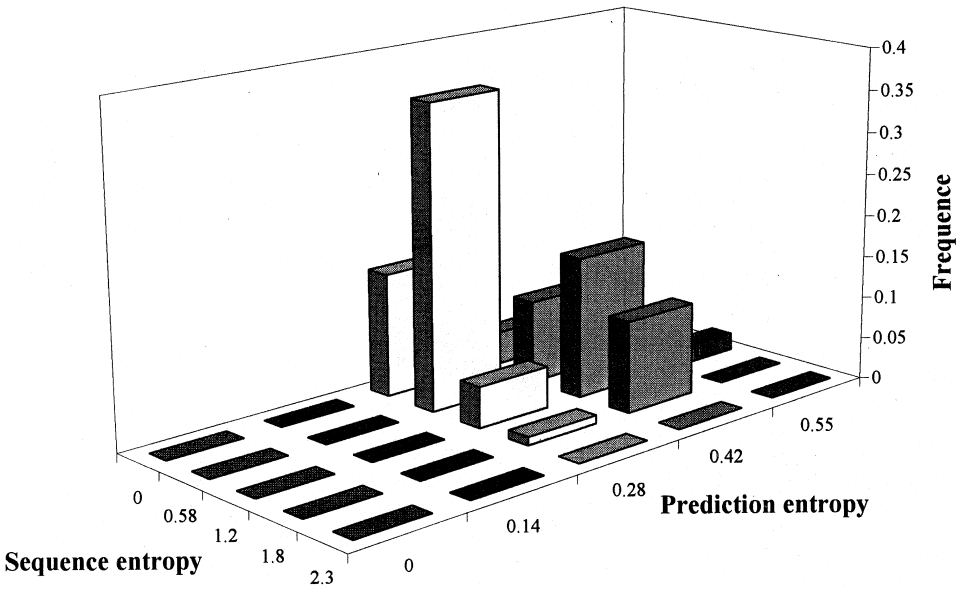

Fig. 6. Three-dimensional histogram of the joint distribution of $S_{\text{seq}}(l)$ and $S_{\text{pred}}(l)$ for the helices of the set $\mathscr{H}^{I}_{100}$ of the homologous proteins examined in Figs. 3 and 5

Table 1 Correlation index ρ of the entropy functions S_{seq} and S_{pred} for sets of residues with different structural properties. N is the cardinality of each set

	$\mathscr{H}^{I}_{100}$	$\mathscr{H}^{nI}_{100}$	$\mathscr{C}_{100}$	$\mathscr{H}^{I}_{50}$	$\mathscr{H}^{nI}_{50}$	$\mathscr{C}_{50}$
ρ	0.410	−0.032	0.026	0.165	−0.031	0.074
N	101	389	755	80	353	668

It is well known that ρ suffers from some ambiguity since although values close to zero mean that linear correlations are virtually absent, they are not sufficient to rule out non-linear functional relationships. The mutual information μ [29] is the appropriate statistic that helps to remove this ambiguity. The values of the mutual information $\mu(\mathscr{C}_k)$ and $\mu(\mathscr{H}^{nI}_k)$ $(k = 50, 100)$ have been calculated by using the probability functions $P(S_{\text{seq}}, S_{\text{pred}})$. The results show them to be very close to zero and this confirms that for the residues lying outside the nucleation helices the entropies $S_{\text{seq}}(l)$ and $S_{\text{pred}}(l)$ are uncorrelated.

4 Conclusions

In conclusion the histograms of Figs. 2 and 3 support the claim that the location of the initiation sites is highly invariant with respect to the sequence identity. Moreover, positional invariance of the nucleation helices has implications for the kinetics of protein folding. Actually in modular theories of folding [11, 30] the location of the early helices constrains the subsequent steps of the folding process and allows estimates of lower bounds of the folding time of the protein. One can thus infer that folding of homologous proteins possessing the same native structure is likely to be characterised by similar intermediate steps. As far as the conservation issue is concerned, there is evidence that sequence variability is reduced in the helical nuclei for which, in addition, the values of $S_{\text{seq}}(l)$ and $S_{\text{pred}}(l)$ are linearly correlated. The distributions of $S_{\text{seq}}(l)$ for the residues not belonging to initiation sites are more similar to a uniform distribution than the plot for the initiation sites (see Figs. 4, 5) and this implies that some of these residues are remarkably well conserved. A reasonable interpretation is that these residues are presumably responsible for some functional feature in the folded protein, although they do not play a key role in the folding process.

Acknowledgements. This work was financially supported by the Ministero per l'Universitá e la Ricerca Scientifica e Tecnologica and by a grant for a Target Project in Biotechnology from the Centro Nazionale per le Ricerche (CNR).

References

1. Anfinsen CB (1973) Science 181:223
2. Fasman GD (Ed) (1990) Prediction of protein structures and the principles of protein conformation. Plenum Press, New York
3. Rost B, Sander C (1993) J Mol Biol 232:584
4. Fariselli P, Compiani M, Casadio R (1993) Eur Biophys J 22:41
5. Casadio R, Fariselli P, Taroni C, Compiani M (1996) Eur Biophys J 24:165
6. Murzin AC, Brenner SE, Hubbard T, Chothia C (1995) J Mol Biol 247:536
7. Compiani M, Fariselli P, Martelli P-L, Casadio R (1998) Proc Natl Acad Sci USA 95:9290
8. Kim PS, Baldwin RL (1990) Annu Rev Biochem 59:631
9. Lesk AM, Rose GD (1981) Proc Natl Acad Sci USA 78:4304
10. Go N (1983) Annu Rev Biophys Bioeng 12:183
11. Karplus M, Weaver DL (1994) Prot Sci 3:650
12. Ptitsyn OB, Uversky YN (1994) FEBS Lett 34:15
13. Abkevich VI, Gutin AM, Shakhnovich EI (1994) Biochemistry 33:10026
14. Wetlaufer DB (1973) Proc Natl Acad Sci USA 70:697
15. Panchenko AR, Luthey-Schulten Z, Wolynes PG (1996) Proc Natl Acad Sci USA 93:2008
16. Panchenko AR, Luthey-Schulten Z, Cole R, Wolynes PG (1997) J Mol Biol 272:95
17. Murphy KP, Bhakuni V, Xie D, Freire E (1992) J Mol Biol 227:293
18. Rooman MJ, Kocher J-P, Wodak SJ (1992) Biochemistry 31:10226

19. Unger R, Moult J (1996) J Mol Biol 259:988
20. Presta LG, Rose GD (1988) Science 240:1632
21. Compiani M, Fariselli P, Casadio R (1997) Phys Rev E 55:7334
22. Chothia C, Lesk AM (1986) EMBO J 5:823
23. Sander C, Schneider R (1991) Proteins 9:56
24. Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res 22:4673
25. Kabsch W, Sander C (1983) Biopolymers 22:2577
26. Khinchin AI (1957) Mathematical foundations of information theory. Dover, New York
27. Shakhnovich E, Abkevich V, Ptitsyn O (1996) Nature 379:96
28. Hoel PG (1984) Introduction to mathematical statistics. Wiley, New York
29. Li W (1990) J Stat Phys 60:823
30. Pappu RV, Weaver DL (1998) Prot Sci 7:480