PERIODICUM BIOLOGORUM
VOL. 101, No 4, 339 - 347, 1999

UDC 57 : 61
CODEN PDBIAD
ISSN 0031-5362

Original scientific paper

Prediction of initiation sites for protein folding with α-helix preferences

DAVOR JURETIĆ[1]
ANA JERONČIĆ[1]
DAMIR ZUCIĆ[2]

[1]Physics Dept.,
Faculty of Natural Sciences,
Univ. of Split,
N. Tesle 12,
21000 Split,
Croatia

[2]Faculty of Electrical Engineering,
Univ. of Osijek,
Istarska 3,
31000 Osijek,
Croatia

Correspondence:
Davor Juretić
Physics, Faculty of Natural Sciences
University of Split
N. Tesle 12
21000 Split,
Croatia

Abstract

Background and purpose: *The formation of α-helices is thought to direct folding in helical and partly helical proteins. Particularly stable helices may be able to retain information about early folding events. Our goal was to test this hypothesis and to develop fast software package for prediction of helix nucleation and folding initiation sites in protein sequence.*

Methods: *A statistical procedure used in this work consists in evaluating folding initiation parameter for each residue in tested sequence by using middle-helix preference functions and geometric average of position specific middle-helix preferences. The best known crystallographic structures of soluble proteins served for the extraction of preference functions. Position specific frequencies of folding initiation sites in observed helices were collected from proteins with known sequence locations of initiation sites.*

Results and conclusions: *The highest frequency of folding initiation sites is in the middle-helix to C-terminus region of experimentally determined helices. Therefore, initiation sites for protein folding are likely to serve as helix-start signals. Overall sequence maximum of our folding initiation parameter is found at the sequence position which belongs to known folding initiation site and to observed α-helix in 84% and 95% of tested sequences respectively. Sequence maximum of this parameter is inside transmembrane helix span for 68% of integral membrane protein sequences of known structure. We developed the Web sever for fast prediction of possible helix nucleation and folding initiation sites at the address: http://pref.etfos.hr/helix-start.*

Key words: prediction, folding initiation, helix nucleation, helix preference, preference functions, web server

INTRODUCTION

When folding begins locally in the sequence it is described as hierarchic folding process *(8, 46)*. Then local elements of regular secondary structure form rapidly and to a good degree persist in the native secondary structure *(35)*. In particular, fast creation of stable α-helices has been observed in peptides *(7, 23, 38, 51)* and in proteins *(35, 40)*. Helix formation is thought to be energetically favored for main-chain atoms of all residues except proline and glycine *(2)*. Still, some residues, such as alanine, have higher helix-forming preference *(12, 13, 18, 38, 42, 43)* and higher helix propagation parameter *s* of the statistical mechanics model for α-helix formation *(10, 53)* than other residues. Also, some position specific steric, electrostatic and hydrophobic interactions are found to favor α-helix formation both in peptides and in proteins *(2, 3, 4, 5, 37)*. The discovery of helix stop signals *(24, 45)* supported the idea that local sequence interactions determine helix nucleation and helix boundaries in native protein structures.

Folding initiation sites have been found in many proteins *(2, 6, 8, 41, 55)*. The prediction of nucleation sites in the sequence where fast helix formation is favored has been recognized as an important goal in folding simulations of helical proteins *(3, 35, 50)*. The hypothesis that α-helices, as local secondary structure, are "seeds for folding" *(45)* has been tested recently *(16)*. It was essential to select reliable helix fragments as good candidates for folding initiation sites. This was achieved by using neural network learning algorithm designed to discriminate between α-helices and non-α structures, and by associating reliable patterns with minimums in the entropy of output vector.

An alternative procedure is to use high helix propensities to select helix fragments of interest. High helix propensities are related to minimums in conformational entropy change, which is an important factor favoring α-helix formation *(17, 20)*. An obvious choice are middle-helix propensities *(34, 49)*. Middle-helix propensities of Kumar and Bansal *(34)* are used in this work for the extraction of corresponding preference functions *(28)* from the database of crystallographic protein structures with the best resolution.

Geometric average of amino acid attributes can also help to predict specific folding motifs *(36)*. Its advantage is that it gives equal weight to high and low attributes in the sequence. In this work we used geometric average of seven position-specific amino acid preferences in the middle-helix positions derived by Richardson and Richardson *(49)*. We report the performance in predicting folding initiation sites in proteins by using the combined index that takes into account both the geometric average of position-specific middle-helix preferences and sequence specific helical preference evaluated with middle-helix preference functions.

In discussion we point out that longer helices, even including transmembrane helices from membrane proteins, are often associated with high maximum for folding initiation index. Locating folding initiation sites in the whole protein and helix-start signals in observed longer helices may help toward goal of selecting helices with a crucial role in early folding history.

TABLE 1

Folding initiation sites and corresponding α-helices. Protein data base considered (PDB codes are used) is the same one that Compiani at al. used (16).

#	protein	INITIATION SITE		corresponding α-helix
		position	segment	position
1)	1rhd	227-236	ELRAMFEAKK	224-235
2)	1phh	236-245	DERFWTELKA	237-245
3)	1cpv	9-20	ADIAAALEACKA	8-17
		63-72	LKLFLQNFKA	60-70
4)	1gd1_o	104-113	DAAKHLEAGA	103-111
		192-201	KDLRRARAAA	-
		257-266	NAALKAAAEG	252-265
5)	1gox	7-15	NEYEAIAKQ	8-16
		142-149	RRAERAGF	134-146
		330-338	MRDEFELTM	309-341
6)	2ts1	4-8	LAELQ	2-10
		279-290	EALEQELREAPE	275-287
7)	3grs	37-44	RRAAELGA	29-43
8)	451c	40-47	AEAELAQR	40-49
9)	2ccy_a	39-50	DAAQRAENMAMV	40-58
		91-98	TESTKLAA	79-102
10)	2cro	3-13	TLSERLKKRRI	3-14
11)	4mdh	160-167	NRAKAQIA	155-171
12)	8adh	333-342	ADFMAKKFAL	323-339
13)	7rsa	2-13	ETAAAKFERQHM	3-13
		25-36	YCNQMMKSRNLT	24-34
14)	1hfx	26-31	WLCIIF	23-34
		89-98	IMCVKKILDI	86-98
15)	2ci2	31-40	SVEEAKKVIL	31-43
16)	2mml	9-17	LVLNVWGKV	4-17
		29-33	LIRLF	21-35
		102-115	KYLEFISECIIQVL	102-118
		133-143	KALELFRKDMA	125-148
17)	1a2p	10-18	VADYLQTYH	7-17
		25-36	ITKSEAQALGWV	27-33
		45-49	VAPG	-
18)	2lza	8-13	LAAAMK	5-14
		28-36	WVCAAKFES	25-35
		92-99	VNCAKKIV	89-100
19)	1hrc	7-15	KKIFVQKCA	3-10
		64-70	LMEYLEN	61-69
		91-101	REDLIAYLKKA	88-101

METHODS

Protein data sets

To extract preference functions *(28)* we used the data set of 100 soluble proteins determined by X-ray analysis (1.7 Å resolution or better) and NMR. There was no more than 30% pairwise sequence identity among these proteins *(54)*. Corresponding Brookhaven Protein Data Bank *(1)*(PDB) codes are listed below:

1aac, 1ads, 1aky, 1amm, 1arb, 1aru, 1ben_a, 1ben_b, 1bkf, 1bpi, 1cem, 1cka, 1cnr, 1cnv, 1cpc_a, 1cpc_b, 1cse_e, 1cse_i, 1ctj, 1cus, 1dad, 1edm, 1fus, 1hfc, 1ifc, 1igd, 1iro, 1isu, 1jbc, 1kap, 1lam, 1lit, 1lkk, 1luc_a, 1luc_b, 1mct_a, 1mct_i, 1mla, 1mrj, 1nfp, 1nif, 1osa, 1phb, 1php, 1plc, 1poa, 1ppn, 1ppt, 1ptf, 1ptx, 1rcf, 1ra9, 1rge, 1rie, 1rro, 1sgp_e, 1sgp_i, 1smd, 1sri, 1snc, 1tca, 1utg, 1vcc, 1whi, 1xic, 1xso, 1xyz, 256b, 2ayh, 2cba, 2cpl, 2ctc, 2end, 2er7, 2erl, 2hft, 2ihl, 2ilk, 2mbw, 2mhr, 2mcm, 2olb, 2phy, 2rhe, 2rn2, 2sil, 2trx, 2wrp, 3b5c, 3chy, 3ebx, 3lzm, 3pte, 3sdh, 4fgf, 4ptp, 5p21, 8abp, 8ruc_a, 8ruc_i.

The data set of 19 proteins with suggested sequence locations of folding initiation sites *(16)* is given in the Table 1 and 2.

The data set of 31 sequences from integral membrane proteins of known crystallographic structure *(30)* is given in the Table 4.

Correlation of folding initiation sites with position of residues in α-helices

The data set of 37 folding initiation sites with corresponding α-helices (Table 1) has been prepared by us based on published data *(9, 14, 22, 25, 26, 27, 41, 47, 50, 52)* for above mentioned 19 proteins. The Compiani et al. *(16)* choice for folding initiation segments is very similar to our choice. Sequence location of secondary structure segments is taken from the most recent PDB assignment. Folding initiation frequencies are calculated as frequencies of residues suggested to initiate folding at specific helix positions.

Middle-helix preferences

Position-specific preferences found by Richardson and Richardson *(49)* for N3, N4, N5, middle, C5, C4 and C3 position of α-helix are here defined as position specific middle-helix preferences. The sliding window with seven amino acid residues scanned the sequence and each heptad score is calculated as the seventh root of the product of seven position-specific preferences. The heptad score is assigned to the middle residue in the scanning window. The result is used directly as the heptad score profile of the sequence and indirectly in the combined parameter (folding initiation parameter, see

TABLE 2

Maximal folding initiation parameters for 19 soluble proteins (see Table 1 legend). Observed secondary structure (SS) is 'H' for α-helix, 'B' for β-sheet, 'U' for undefined, coil or turn structure and 'F' for folding initiation site (usually in the α-helix conformation). The heptad segments are centered at amino acid number (AA) with sequence maximum for the combined index (FIP). Heptad maximums are found with geometric average of middle-helix preferences (49), *while preference maximums are found by evaluating preference functions extracted with Kumar and Bansal* (34) *middle-helix preferences.*

Protein	AA	heptade max	SS	AA	preference max	SS	AA	combined index max	SS	heptad segment
1rhd	230	2.178	F	232	2.163	F	230	3.666	F	ELRAMFE
1phh	111	1.510	H	334	2.294	H	334	3.452	H	ICLRRIW
1cpv	65	1.826	F	15	2.650	F	14	3.744	F	IAAALEA
1gd1_o	258	1.792	F	199	2.540	F	261	3.810	F	AALKAAA
1gox	297	1.547	H	368	2.772	U	142	3.641	F	QLVRRAE
2ts1	169	1.868	H	418	2.634	U	314	3.836	H	ALRQAIR
3grs	449	1.401	H	2	2.584	U	39	3.595	F	ARRAAEL
451c	46	1.772	F	43	2.559	F	45	3.667	F	AELAQRI
2ccy_a	98	1.601	H	45	2.647	F	46	3.770	F	RAENMAM
2cro	52	1.477	H	71	2.806	U	12	3.615	F	KKRRIAL
4mdh	99	1.660	U	165	2.466	F	165	3.691	F	AKAQIAL
8adh	38	1.671	U	11	2.086	U	338	3.345	F	FMAKKFA
7rsa	28	1.657	F	6	2.294	F	6	3.643	F	TAAAKFE
1hfx	28	1.622	F	123	2.926	U	119	3.180	U	EQWYCFA
2ci2	48	1.546	B	38	1.841	F	37	3.116	F	EAKKVIL
2mm1	47	1.600	U	133	2.267	F	133	3.428	F	AMNKALE
1a2p	16	1.572	F	1	2.512	U	32	2.853	F	EAQALGW
2lza	10	2.188	F	9	2.409	F	10	3.810	F	ELAAAMK
1hrc	96	1.676	F	97	2.116	F	96	3.420	F	DLIAYLK

below).

Middle-helix preferences of Kumar and Bansal *(34)* are not position specific. We used a single set of 20 preferences that Kumar and Bansal extracted from the middle section of helices longer than eight residues. The data set of 100 soluble proteins served to extract corresponding middle-helix preference functions *(28, 29)*. The evaluation of preference functions in tested sequence produced sequence dependent conformational preferences *(28)*. In the combined index heptad scores are smoothed as three point averages and added to sequence dependent helical preference. In the following text the term combined index or folding initiation parameter (FIP) is used for the

TABLE3

Prediction accuracy for folding initiation sites and for α-helices longer than eight residues.

Protein data base:	19 Proteins		100 proteins
Prediction accuracy* (%)	Folding initiation sites	α-helices	α-helices
Per-segment			
Sensitivity	84	73	63
Efficiency	41	61	51
Per-residue			
Sensitivity	68	41	31
Efficiency	24	73	63

* for the FIP threshold (see text) ≥ 2.6

TABLE 4

Maximal heptad score and maximal folding initiation parameter (FIP) values for 31 sequences of membrane polypeptides with known sequence location of transmembrane helices (TMH). The letter "Y" in the last column denotes the case when the FIP sequence maximum is find inside the TMH span, while N denotes the case when this is not so. N (C) is the case when maximum is found at the protein C-terminus.

protein	AA	maximal heptade score	SS	AA	maximal combined index (FIP)	SS	TMH
lprc_h	23	1.422	H	17	3.483	H	Y
laig_h	176	1.843	U	256	3.822	U	N(C)
lprc_l	264	1.636	H	103	3.094	H	Y
laig_l	264	1.677	H	126	3.593	H	Y
lprc_m	185	1.570	H	246	3.586	H	N
laig_m	216	1.541	H	248	3.758	H	N
lkzu_a	27	1.284	H	31	3.121	H	Y
P04159	153	1.671	H	100	3.377	H	Y
larl_a	99	1.760	H	339	3.741	H	Y
larl_b	64	1.729	H	169	3.624	U	N
P06030	107	1.722	H	106	3.338	H	Y
locc_a	290	1.785	H	468	3.604	H	Y
locc_b	150	1.550	B	74	3.155	H	Y
locc_c	225	1.733	U	162	3.555	H	Y
locc_d	84	1.488	H	41	3.404	H	N
locc_g	16	1.332	H	17	2.406	H	Y
locc_i	62	1.462	H	15	3.404	H	Y
locc_j	10	1.516	H	9	3.385	H	N
locc_k	30	1.304	H	30	2.972	H	Y
locc_l	26	1.512	H	22	3.653	H	Y
locc_m	34	1.355	H	1	2.475	U	N
lbcc_e	70	1.528	U	111	3.502	H	N
lbe3_k	15	1.420	U	14	2.764	U	N
lbcc_j	53	1.602	H	32	3.145	H	Y
lbcc_g	55	1.386	H	43	3.250	H	Y
lbcc_d	63	1.633	H	63	3.684	H	N
lbe3_c	315	1.568	H	238	3.969	H	Y
lbrx	12	1.699	H	146	3.605	H	Y
lafo	4	1.567	U	36	2.948	H	Y
lbl8	90	1.474	H	7	3.382	H	Y
la91	13	1.442	H	13	3.786	H	Y

* PDB codes except for the Swiss-Prot codes P04159 and P06030.

combination of Richardson & Richardson and Kumar & Bansal preferences as described above.

The sequence location of transmembrane helices is predicted by using the SPLIT 3.5 suite of algorithms *(29, 30)* available at our Web server: http://pref.etfos.hr/split. Computer program for the calculation of folding initiation parameters was written in FORTRAN 77. It was translated into ANSI C and wrapped into the Web server HELIX-START, written in HTML and Unix shell script language. A graphic library, created for the HELIX-START server, enables fast (in seconds) graphical presentation of calculated profiles by using the server at the address: http://pref.etfos.hr/helix-start.

Prediction accuracy calculations

Prediction accuracy is reported as per-segment and per-residue accuracy. Only α-helical segments longer than eight residues are considered for segment prediction of helices, but all residues observed in the helical conformation are considered for the per-residue prediction of α-helix conformation. Prediction accuracy for folding initiation sites is calculated for the data set of 19 proteins in which such sites are known.

To take into account overpredictions we report prediction accuracy as efficiency (# correct predictions/ # predictions), and as sensitivity (# correct predictions/ # observed features). Correct per-residue prediction is scored whenever higher then threshold FIP value is found inside observed feature. Overprediction is scored when higher then threshold value is found outside experimentally determined folding motifs that are being predicted. Per-segment prediction accuracy is calculated by taking into account only maximal FIP value inside observed segment. Higher than threshold FIP value, found during sequence scan outside observed segments, is scored as segment overprediction, if corresponding residue is a) at least three residues removed from the C-terminus end of observed segment, and b) at least eight residues removed from the residue already scored for segment overprediction. First and last eight residues in a protein are not considered for segment overprediction.

RESULTS

Prediction of folding initiation sites in soluble proteins

Folding initiation sites in proteins are often found in the α-helix structure *(16, 50)*, which is not surprising because α-helices are secondary structure elements known to fold very fast *(35)*. Are folding initiation sites in helical proteins particularly strong helix-start signals, and if so are they found more often closer to helix N-terminus, helix-middle or helix C-terminus? The analysis of protein data set with known or suggested strong folding initiation sites revealed that frequencies of occurrence of folding initiation residues are maximal at middle-helix positions with a slight preference for the C-terminal half of the helix (Figure 1).

In coiled coil helices *(15)* helix-start signals are likely to be hidden in each heptad of amino acid residues. If so, can we use a technique similar to Lupas et al. *(36)* to associate heptades having high score for middle-helix position with strong helix-start signals and potential folding initiation sites ? We used seven columns of position specific middle-helix preferences *(49)* to find such scores as described in the Methods section. We found that almost all folding initiation sites are indeed associated with at least one high heptad score, while highest heptad score for the whole sequence is often associated with the folding initiation site (columns 2, 3 and 4 in the Table 2).

Since chosen Richardson & Richardson preferences were for middle helix positions we asked if preference functions derived from middle helix preferences *(29, 30)* are as good indicators of folding initiation sites as heptad scores. For 11 out of 19 proteins the highest helical preference in the whole sequence, evaluated with Richardson & Richardson preference functions *(29, 30)*, is found to be located inside folding initiation site. Out of 37 folding initiation segments in these proteins all but one are associated with a maximum in the sequence dependent α-helix preference (not shown).

Are these results reproducible with middle-helix preferences other than Richardson & Richardson's? To answer this question we used middle-helix preferences of Kumar and Bansal *(34)*. Corresponding preference functions were equally good predictors of folding initiation sites (columns 5, 6 and 7 in the Table 2).

Is combined index, defined as the sum of heptad score and sequence dependent α-helix preference (evaluated with middle-helix preference functions) superior to above mentioned predictors? Sequence maximum in the combined index is associated with folding initiation site for 14 out of 19 sequences when middle-helix preferences are evaluated with Richardson & Richardson preference functions *(29)*. Even better

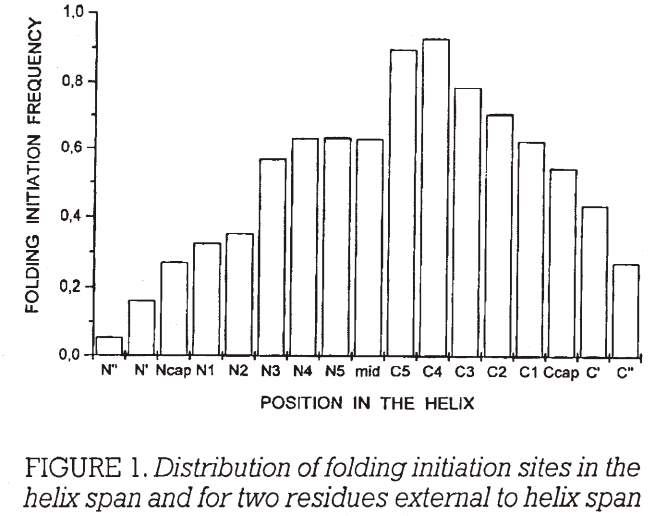

FIGURE 1. *Distribution of folding initiation sites in the helix span and for two residues external to helix span (N", N', and C', C"). Folding initiation frequencies were calculated as frequencies of folding initiation sites at specific helix positions.*

result for the combined index (Table 2) is achieved when middle-helix preferences are evaluated with Kumar and Bansal preference functions. Out of 19 sequence maximums for the combined index 16 are associated with folding initiation sites (Table 2) and 18 are associated with the observed α-helix structure. Therefore, the term folding initiation parameter (FIP) seems to be appropriate for the combined index evaluated in this manner.

What would be a good choice for the threshold FIP value suitable for predicting folding initiation sites? With FIP ≥ 2.6 we find 230 out of 339 folding initiation residues and 31 out of 37 folding initiation segments listed in the Table 1. Segments are found by looking if maximal FIP value in the segment is greater or equal to 2.6. Lower FIP threshold increases the percentage of correct predictions (36 folding initiation segments are predicted with a threshold of 2.0), but produces too many overpredictions. It is possible of course that overpredictions of folding initiation sites are in fact correct predictions for helix-start signals that would uncover sequence location of some native helices.

Prediction of the α-helix conformation with the FIP index

For the data set of 19 proteins where sequence location of α-helices and folding initiation sites are both known it is possible to predict both features and to compare the prediction accuracy (Table 3). Correct prediction (prediction sensitivity) of 88% longer α-helices is achieved with the FIP threshold greater or equal to 2.0. Even higher sensitivity of 97% for predicting folding initiation segments is accompanied with low prediction efficiency of 21%. Similar high percentage (82%) of correct prediction of longer helices is found for the data set of 100 soluble proteins (Methods) with the same FIP threshold. With increasing FIP threshold the number of correct predictions drops and becomes similar for folding initiation segments and for longer helices (Figure 2). This is not accidental. Helices that are not a part of the initiation sites are eliminated by setting a high threshold. For instance in the case of the highest threshold considered (3.6) a total of 12 correctly predicted folding initiation sites corresponds to 11 out of 12 correctly predicted helices, while only one helix is overpredicted (prediction efficiency of 92%).

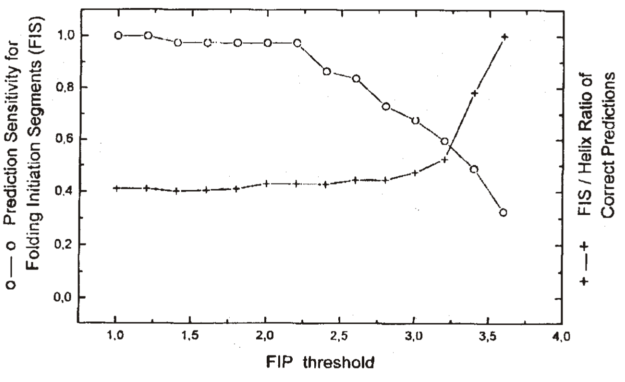

FIGURE 2. *The dependence of the prediction accuracy on the chosen threshold for the folding initiation parameter (FIP). The ratio of correctly predicted to observed folding initiation segments (FIS, open circles) decreases, while the ratio of correctly predicted FIS to helical segments (plus symbols) increases with the increase in the FIP threshold.*

Majority of helical residues are underpredicted. For instance in the case of 5288 helical residues out of 18769 in the data set of 100 soluble proteins the FIP value of 2.6 or greater correctly predicts 1635, overpredicts 953 and underpredicts 3653. Underprediction is less serious for initiation sites. Out of 339 such sites (residues) in 19 proteins 109 are underpredicted with the same FIP threshold.

Folding initiation sites in membrane proteins

For integral membrane proteins overall sequence maximum for folding initiation parameter is found inside observed transmembrane helix in 68% of sequences (Table 4). The percentage raises to 77% when remaining sequences are examined for sequence maximum of heptad scores. When extramembrane helices are taken into account as well then 87% of membrane protein sequences are found with maximal FIP inside some helix. For 31 membrane proteins (Table 4) 57% of residues are associated with the α-helix conformation.

Examples of profiles for folding initiation parameter

The profile of folding initiation parameter is shown for the cytochrome c (1hrc) (Figure 3). Three high maximums correspond to known folding initiation segments *(26)* (shaded columns for amino acids 7-15, 64-70 and

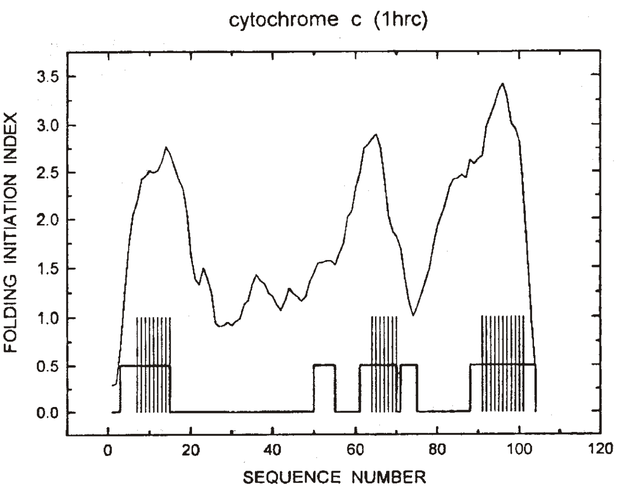

FIGURE 3. *The profile of folding initiation parameter for the cytochrome c (1hrc). Known folding initiation sites are shown as shaded columns up to the height of 1.0. Observed α-helices are shown as the bold line at the 0.5 level.*

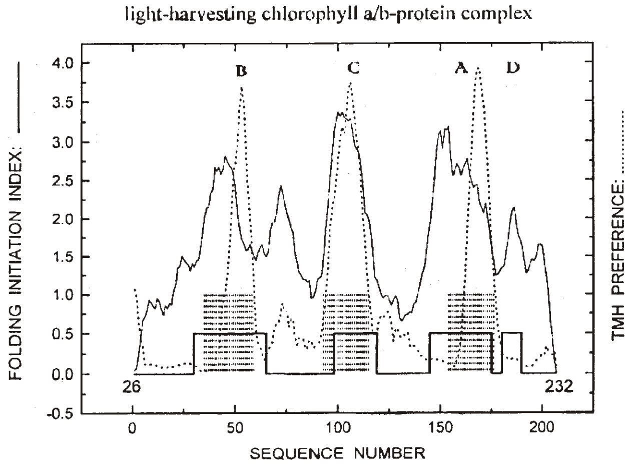

FIGURE 4. *The profile of folding initiation parameter (full thin line) and transmembrane helix preference (dotted line) for light-harvesting protein. Transmembrane helical preferences are evaluated from corresponding preference functions* (29). *The observed span of transmembrane helices A, B, C and of surface helix D is shown as the bold line at the 0.5 level. Predicted locations of transmembrane helices with the web server SPLIT* (30) *are shown as shaded rows up to the 1.0 level. Reported profiles are for the protein fragment 26-232*

91-101) and to observed longer helices (bold line at the 0.5 level for amino acids Val 3 - Cys 14, Glu 61 - Glu 69 and Lys 88 - Asn 103). Hydrophobic-hydrophobic contacts with heme ligand are also found at these sequence positions (amino acids 10, 13, 14, 64, 67, 68, 94, 98)

Another example of the FIP profile (Figure 4) is given for the apoprotein of the major light-harvesting complex of photosystem II in plant (*Pisum sativum*) *(33)*. It has three FIP maximums that are almost as good indicators of the position of observed membrane spanning helices as Kyte-Doolittle preference functions (dotted line) *(28, 29)*. It is also of interest that folding initiation maximums are associated with chlorophyll side chain ligands Glu 65, His 68 (helix B), Gln 131, Glu 139 (helix C) and Glu 180, Asn 183 (helix A) (notice that correct sequence numbers on the x-axis are obtained by addition of 25 N-terminal amino acids, omitted in the reported structure *(33)*).

DISCUSSION

Computational methods of sequence analysis with a goal to model folding process can profit from evidence that helix formation can direct folding. Presta and Rose *(45)* proposed that clusters of residues with high helix preference at the helix boundaries are necessary for helix formation during protein folding. Helix-start signals are expected to occur closer to the N-terminal helix positions *(11)*, while helix-stop signals are often found at the helix N-terminus *(24)*. However, helix-start signals, that can serve as folding initiation sites as well, are not located predominantly at the helix N-terminus (Figure 1). Folding initiation sites in 19 proteins that we considered are best associated with middle to C-terminus helical region. The frequency of folding initiation drops toward helix C-terminus and beyond, but it is still significantly higher than corresponding frequency for the N-terminal part of helix. Therefore, at least for one class of folding initiation sites in soluble proteins, helix nucleation of specific native helices initiates folding at or close to nascent middle-helix region

Instead of taking into account all possible position-specific local interactions favoring helix formation we used position specific middle-helix preferences of Richardson and Richardson *(49)* and preference functions *(28)* calculated with middle-helix preferences of Kumar and Bansal *(34)*. The justification for such a procedure for predicting folding initiation sites is a) calculated frequency of folding initiation sites which is maximal close to middle-helix positions (Figure 1) and b) expectation that helix formation dominates the folding kinetics of helical protein *(35)*.

Our procedure, which uses middle-helix preferences to calculate sequence profile of folding initiation index, may seem complicated, but the interpretation of the profile is straightforward. By choosing a high FIP threshold (3.6) one can identify those nascent helices that are crucial for initiation of protein folding. Prediction efficiency for finding such helices is higher than 90% when tested at limited data set of 19 soluble proteins with known folding initiation sites. About one third of observed folding initiation segments are then found with very few false-positive predictions. With a lower threshold (2.6) more than 80% of observed initiation segments are found and more than 70% of observed longer helices (Table 3). Still lower FIP threshold would produce even better prediction sensitivity, but decreased efficiency. For comparison the sensitivity of 34% was reported in predicting nucleation of protein helices with strip-of-helix hydrophobicity algorithm *(48)*.

Hem and chlorophyll ligands are known to promote helix formation *(32, 44)*. Therefore, it is not surprising that amino acid contacts with such ligands in two sequences we examined are found to be associated with high folding initiation potential.

Sequence span of observed transmembrane helices in integral membrane proteins is underpredicted when prediction is based on hydrophobicity analysis *(29)*. While hydrophobicity, or helix preference based on hydrophobicity, is as a rule maximal in the middle region of membrane spanning helix this is not so for FIP maximums. FIP maximums are generally found closer to N or C-terminus of transmembrane helices (Figure 4 and unpublished observations). Richardson & Richardson preference functions were used by us recently to refine the prediction of transmembrane helices and for the prediction of interface helices in membrane proteins *(29, 30)*. The FIP index too has the potential to improve the prediction of transmembrane helix boundaries.

Postulated folding mechanism of rapid hydrophobic collapse *(39)* has to be braked in those membrane proteins, whose entrance in the membrane requires partially unfolded structure *(19, 21)*. The separation of early folding and hydrophobic domain in some of future membrane spanning domains is probably important for the membrane entry process. The light-harvesting protein enters thylakoid membrane from the stromal space so that hydrophobic C-terminals of helices A and B are oriented toward thylakoid space, while early folding N-terminal parts of these helices are oriented toward the stroma. If membrane entry is mediated by the translocase complex *(31)* than many charges in the N-terminal parts of helices A and B may assume specific configuration in the α-helix conformation early in the folding history facilitating specific interactions with chlorophylls and with the translocation apparatus.

In conclusion, for soluble helical or partly helical proteins, the initiation sites for protein folding correspond to sequence regions with strong middle-helix preference. By setting a high threshold for our FIP parameter one can select helices that initiate protein folding. In membrane proteins maximal FIP values are often associated with interface regions of transmembrane helices and can reveal topological signals (work in progress).

Acknowledgements: We thank Bono Lučić from Ruđer Bošković Institute in Zagreb who supplied us with copies of some papers needed for this work. Croatian Ministry of Science supported this research with the grant 177060. Thanks are also due to thoughtful comments by the anonymous reviewer.

REFERENCES

1. ABOLA E, BERNSTEIN F C, BRYANT S H, KOETZLE T F, WENG J 1987 Protein Data Bank. *In:* Crystallographic Databases-Information Content, Software Systems, Scientific Applications, ALLEN F H, BERGERHOFF G, SIEVERS R *(eds)*, p 107-132. Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester.

2. AURORA R, CREAMER T P, SRINIVASAN R, ROSE G D 1997 Local interactions in protein folding: lessons from the α-helix. *J Biol Chem 272:* 1413-1416

3. AVBELJ F, MOULT J 1995a The conformation of folding initiation sites in proteins determined by computer simulation. *Proteins: Struct Funct Genet 23:* 129-141

4. AVBELJ F, MOULT J 1995b Role of electrostatic screeining in determining protein main chain conformational preferences. *Biochemistry 34:* 755-764

5. AVBELJ F, FELE L 1998 Role of main-chain electrostatics, hydrophobic effect and side-chain conformational entropy in determining the secondary structure of proteins. *J Mol Biol 279:* 665-684

6. BALDWIN R L 1986 Seeding protein folding. *Trends in Biochemical Sciences 11:* 6-9

7. BALDWIN R L 1995 α-Helix formation by peptides of defined sequence. *Biophys Chem 55:* 127-135

8. BALDWIN R L, ROSE G D 1999 Is protein folding hierarchic? I. Local structure and peptide folding. *Trends in Biochemical Sciences 24:* 26-33

9. BYCROFT M, MATOUSCHEK A, KELLIS J T JR, SERRANO L, FERSHT A R 1990 Detection and characterization of a folding intermediate in barnase by NMR. *Nature 346:* 488-490

10. CHAKRABARTTY A, SCHELLMAN J A, BALDWIN R L 1991 Large differences in the helix propensities of alanine and glycine. *Nature 351:* 586-588

11. CHAKRABARTTY A, DOIG A J, BALDWIN R L 1993 Helix capping propensities in peptides parallel those in proteins. *Proc Natl Acad Sci U S A 90:* 11332-11336

12. CHOU P Y, FASMAN G D 1974a Conformational parameters of amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry 13:* 211-222

13. CHOU P Y, FASMAN G D 1974b Prediction of protein conformation. *Biochemistry 13:* 222-245

14. CHYAN C L, WORMALD C, DOBSON C M, EVANS P A, BAUM J 1993 Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: a hydrogen exchange study. *Biochemistry 32 :* 5681-5691

15. COHEN C, PARRY DAD 1990 α-Helical coiled coils and bundles: How to design a α-helical protein. *Proteins Struct Funct Genet 7:* 1-15

16. COMPIANI M, FARISELLI P, MARTELLI PL, CASADIO R 1998 An entropy criterion to detect minimally frustrated intermediates in native proteins. *Proc Natl Acad Sci USA 95:* 9290-9294

17. CREAMER T P, ROSE G D 1992 Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc Natl Acad Sci U S A 89:* 5937-5941

18. CREAMER T P, ROSE G D 1994 α-Helix forming propensities in peptides and proteins. *Proteins Struct Funct Genet 19:* 85-97

19. DALBEY R E, ROBINSON C 1999 Protein translocation into and across the bacterial plasma membrane and the plant thylakoid membrane. *Trends in Biochemical Sciences 24:* 17-22

20. DOIG A J, STERNBERG M J E 1995 Side-chain conformational entropy in protein folding. *Protein Sci 4:* 2247-2251

21. EILERS M, SCHATZ G 1986 Binding of a specific ligand inhibits import of a purified precursor protein. *Nature 322:* 228-232

22. FERSHT A R 1995 Optimization of rates of protein folding: The nucleation-condensation mechanism and its implications. *Proc Natl Acad Sci USA 92:* 10869-10873

23. GRUENEWALD B, NICOLA C U, LUSTIG A, SCHWARZ G 1979 Kinetics of the helix-coil transition of a polypeptide with non-ionic side groups, derived from ultrasonic relaxation measurements. *Biophys Chem 9:* 137-147

24. HARPER E T, ROSE G D 1993 Helix stop signals in proteins and peptides: The capping box. *Biochemistry 30:* 7605-7609

25. HUGHSON F M, WRIGHT P E, BALDWIN R L 1990 Structural characterization of a partly folded apomyoglobin intermediate. *Science 249 :*1544-1548

26. JENG M F, ENGLANDER S W, ELOVE G A, WAND A J, RODER H 1990 Structural description of acid-denatured cytochrome c by hydrogen exchange and 2D NMR. *Biochemistry 29:* 10433-10437

27. JENNINGS P A, WRIGHT P E 1993 Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science 262:* 892-896

28. JURETIĆ D, ZUČIĆ D, LUČIĆ B, TRINAJSTIĆ N 1998a Preference functions for prediction of membrane-buried helices in integral membrane proteins. *Computers Chem. 22:* 279-294

29. JURETIĆ D, LUČIN A 1998b The preference functions method for predicting protein helical turns with membrane propensity. *Journal of Chemical Information and Computer Sciences 38:* 575-585

30. JURETIĆ D, JERONČIĆ A, ZUČIĆ D 1999 Sequence analysis of membrane proteins with the web server SPLIT. *Croatica Chemica Acta (in press)*

31. KIM S J, JANSSON S, HOFFMAN N E, ROBINSON C, MANT A 1999 Distinct "assisted" and "spontaneous" mechanisms for the insertion of polytopic chlorophyll-binding proteins into the thylakoid membrane. *J Biol Chem 274:* 4715-4721

32. KRIGBAUM W R, KNUTTON S P 1973 Prediction of the amount of secondary structure in a globular protein from its amino acid composition. *Proc Natl Acad Sci USA 70:* 2809-2813

33. KÜHLBRANDT W, WANG D N, FUJIYOSHI Y 1994 Atomic model of plant light-harvesting complex by electron crystallography. *Nature 367:* 614-621

34. KUMAR S, BANSAL M 1998 Geometrical and sequence characteristics of α-helices in globular proteins. *Biophysical Journal 75:* 1935-1944

35. LAURENTS D V, BALDWIN R L 1998 Protein folding: matching theory and experiment. *Biophys J 75:* 428-434

36. LUPAS A, VAN DYKE M, STOCK J 1991 Predicting coiled coils from protein sequences. *Science 252:* 1162-1164

37. LOCKHART D J, KIM P S 1993 Electrostatic screening of charge and dipole interactions with the helix backbone. *Science 260:* 198-202

38. MARQUSEE S, ROBBINS V H, BALDWIN R L 1989 Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci USA 86:* 5286-5290

39. MATHESON R R, SCHERAGA H A 1978 A method for predicting nucleation sites for protein folding based on hydrophobic contacts. *Macromolecules 11:* 819-829

40. MATTHEWS C R 1993 Pathways of protein folding. *Annu Rev Biochem 62:* 653-683

41. MOULT J, UNGER R 1991 An analysis of protein folding pathways. *Biochemistry 30:* 3816-3824

42. O'NEIL K T, DeGRADO W F 1990 A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science 250:* 646-651

43. PACE C N, SCHOLTZ J M 1998 A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J 75:* 422-427

44. PAULSEN H, FINKENZELLER B, KUHLEIN N 1993 Pigments induce folding of light-harvesting chlorophyll a/b-binding protein. *Eur J Biochem 215:* 809-816

45. PRESTA L G, ROSE G D 1988 Helix signals in proteins. *Science 240:* 1632-1641

46. PTITSYN O B 1998 Protein folding: nucleation and compact intermediates. *Biochemistry 63:* 367-373

47. RADFORD S E, DOBSON C M, EVANS P A 1992 The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature 358:* 302-307

48. REYES V E, PHILLIPS L, HUMPHREYS R E, LEW R A 1989 Prediction of protein helices with a derivative of the strip-of-helix hydrophobicity algorithm. *J Biol Chem 264:* 12854-12858

49. RICHARDSON J S, RICHARDSON D C 1988 Amino acid preferences for specific locations at the ends of a helices. *Science 240:* 1648-1652

50. ROOMAN M J, KOCHER J-P A, WODAK S J 1992 Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry 31:* 10226-10238

51. SCHOLTZ J M, BALDWIN R L 1992 The mechanism of α-helix formation by peptides. *Annu Rev Biophys Biomol Struct 21:* 95-118

52. UDGAONKAR J B, BALDWIN R L 1990 An early folding intermediate of ribonuclease A. *Proc Natl Acad Sci USA 87:* 8197-8201

53. WOJCIK J, ALTMANN K H, SCHERAGA H A 1990 Helix-coil stability constants for the naturally occurring amino acids in water. XXIV Half-cystine parameters from random poly(hydroxybutylglutamine-co-S-methylthio-L-cysteine. *Biopolymers 30:* 121-134

54. WORD J M, LOVELL S C, RICHARDSON J S, RICHARDSON D C 1999 Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol 285:* 1735-1747

55. WRIGHT P E, DYSEN J, LERNER R A 1988 Conformation of peptide fragments of proteins in aqueous solution: Implications for initiation of protein folding. *Biochemistry 27:* 7167-7175