A Data Base of Minimally Frustrated Alpha Helical Segments Extracted from Proteins According to an Entropy Criterion

Rita Casadio [a, b], **Mario Compiani** [a, c], **Piero Fariselli** [a] **and Pier Luigi Martelli** [a]

(a) CIRB, Biocomputing unit, University of Bologna; (b) Laboratory of Biophysics, Dept.of Biology, Via Irnerio 42, I-40126-Bologna, Italy; (c) Dept of Chemical Sciences, University of Camerino; casadio@kaiser.alma.unibo.it, compiani@camserv.unicam.it, (piero, gigi)@lipid.biocomp.unibo.it

Abstract

A data base of minimally frustrated alpha helical segments is defined by filtering a set comprising 822 non redundant proteins, which contain 4783 alpha helical structures. The data base definition is performed using a neural network-based alpha helix predictor, whose outputs are rated according to an entropy criterion. A comparison with the presently available experimental results indicates that a subset of the data base contains the initiation sites of protein folding experimentally detected and also protein fragments which fold into stable isolated alpha helices. This suggests the usage of the data base (and/or of the predictor) to highlight patterns which govern the stability of alpha helices in proteins and the helical behavior of isolated protein fragments.

Keywords

Alpha helix prediction; neural networks; minimal entropy criterion; protein folding; self-stabilizing alpha helices.

Introduction

Proteins can be considered frustrated systems, namely systems for which the simultaneous minimization of all interaction energies is impossible (Bryngelson and Wolynes, 1987; Frauenfelder and Wolynes, 1994; Bryngelson et al., 1995). However, the tendency of short amino acid sequences to adopt stable conformational states in solution (referred to as minimally frustrated segments) is believed to drive protein folding towards the native structure. This notion is supported by theoretical analysis of protein folding (Rooman et al., 1992; Sali et al., 1994; Abkevich et al., 1994; Karplus and Weaver, 1994; Dill and Chan, 1997; Klimov and Thirumulai, 1998) and by experimental results, which indicate that one of the earliest steps in the folding process is the formation of secondary structure elements, essentially stabilized by short range interactions, which may or may not be concomitant with a hydrophobic collapse (Serrano et al., 1992; Matthews, 1993; Qian and Chan, 1996; Hao and Scheraga, 1998). Alpha helix structures, which are locally stabilized by intrachain interactions, are viable candidates for the role of minimally frustrated folding initiation sites (Presta and Rose, 1988) and may be regarded as a suitable simplified model of protein folding (Muñoz and Serrano, 1995). Moreover a large body of information is available on the factors contributing to the stability of alpha helices in proteins and to the helical behavior of protein fragments in solution (Scholtz and Baldwin, 1992; Muñoz and Serrano, 1994; Baldwin, 1995 and references therein).

In this paper we analyze a data set of 822 non redundant proteins, containing 4783 alpha helix structures with the aim of extracting the minimally frustrated alpha helical fragments and defining a data base containing the most stable alpha helical fragments in proteins of known structure. This is performed using a neural network-based alpha helix predictor which implements a minimum entropy criterion for the identification of the minimally frustrated alpha helical fragments (Compiani et al., 1998). Comparison with experimental results clearly indicates that protein folding initiation sites and stable isolated alpha helical protein fragments are a subset of our data base.

The protein training set

A set of 822 proteins was extracted from the non homologous protein data base (with an identity value <25%) selected using the PDB_select_jun_98 algorithm (http://www.embl-heidelberg.de). This set comprises 174191 amino acid residues, 52618 of which are classified alpha helix by the DSSP program (Kabsch and Sander, 1983) for a total of 4783 alpha helix structures. Following the definitions suggested by Zhang and Chou (1992), 137 proteins of the set are all alpha, 37 all beta, and 114 alpha/beta.

Table 1: The prediction efficiency of the network with single sequence and multiple sequence inputs

Single Sequence			
Q_2 = 0.70	Q(H) = 0.81	Q(nonH) = 0.65	Sov_H^{pred} = 0.84
C = 0.43	Pc(H) = 0.50	Pc(nonH) = 0.89	Sov_H^{obs} = 0.58
Multiple Sequence			
Q_2 = 0.85	Q(H) = 0.67	Q(nonH) = 0.93	Sov_H^{pred} = 0.85
C = 0.63	Pc(H) = 0.80	Pc(nonH) = 0.86	Sov_H^{obs} = 0.76

Q_2= accuracy of the prediction in 2 states (helix, non-helix). C=Correlation coefficient. Q(H)=correctly predicted/observed residues for the helical state. Q(nonH)=correctly predicted/observed residues for the non helical state. Pc(H)=correctly predicted/total predicted residues for the helical state. Pc(nonH)= correctly predicted/total predicted residues for the non helical state. Sov_H^{pred} , Sov_H^{obs} = predicted and observed measures of the overlapping segments as defined in Rost and Sander (1994).

The neural network-based predictor and the entropy criterion

A standard feed-forward neural network is implemented with a back propagation algorithm as learning procedure (Rumelhart et al., 1986). The network architecture consists of a perceptron with one hidden layer containing 22 hidden nodes and an input window spanning 17 residues. Two output nodes discriminate alpha helices from other structures, respectively. Evolutionary information is used as input (Rost and Sander, 1994) to improve the performance of the two-state discriminating network using the single sequence input (Table 1). Proteins are presented to the network using their sequence profile, as taken from the HSSP files (Sander and Schneider, 1991). Each residue is encoded by a 20-element vector, where each element represents the frequency of the residue in the multiple sequence alignment. A cross validation procedure is used to validate the predictor efficiency. This is done by splitting the whole set of proteins into 20 subsets containing an approximate equal number of chains. One subset at a time is removed from the training set and used as testing set. For the evaluation of the statistical indices which score the predictor efficiency (Table 1), the predictions of the 20 different networks are averaged. Under these conditions the accuracy of the prediction (Q_2) is 0.85. To our knowledge, this accuracy value is higher than those of the available competing systems performing the same task (Rost and Sander, 1993; Mamitsuka and Yamanishi, 1995). Moreover, a neural network-based predictor trained with the multiple sequence profile to discriminate three structural states (alpha, beta and coil), similar to that known to perform with the highest efficiency when compared to the other three-state discriminating predictors available to date (Cuff and Barton, 1999), was also tested on the two-state prediction of our data base (merging beta and coil outputs). In this case, the accuracy of the predictor (Q_2), which is 0.72 when in cross validation three structural types are discriminated, becomes 0.83. This value is somewhat lower than that of the two-state discriminating predictor (described above), which is used in this study also in order to make possible the probabilistic interpretation of the outputs. A standard way to assess the network's output is based on the reliability index (Rost and Sander, 1994). However we use the more statistically grounded information entropy (Eq.1) which turns out to be a more sensitive measure in the low entropy (high reliability index) range (Fig.1).

The network outputs are rated according to their entropy value, which is computed using the following equation:

$$S = -\Sigma\, p(i)ln(p(i)) \qquad (1)$$

Considering that the network outputs *p(i)* are real numbers in [0 , 1], the entropy (*S*) limiting values are 0 and 0.693. This procedure is based on the theorem which ensures the convergence of the outputs of the back propagation networks to the conditional probability that a given input is found in a given structure type (Bishop, 1994). In order to regularize the entropy profile of the predicted chain, the entropy value of each predicted residue is averaged over a segment comprising 5 contiguous residues (S5). The segments characterized by a minimum entropy value in the chain correspond to the most reliable patterns predicted by the network and also to the minimally frustrated fragments in the protein. This was previously demonstrated (Compiani et al., 1998), considering that the structural assignment carried out by the network on the basis of the input pattern relies on an average over all the contexts of the training set. The averaging procedure is the main performance limiting factor for those patterns whose secondary structure is protein-context dependent (i.e., those segments which are stabilized by long range tertiary interactions in the protein); conversely, it scarcely affects the prediction of those patterns whose secondary structure is largely context-independent (i.e., those segments referred to as minimally frustrated fragments which are

stabilized by short range interactions in the protein context).

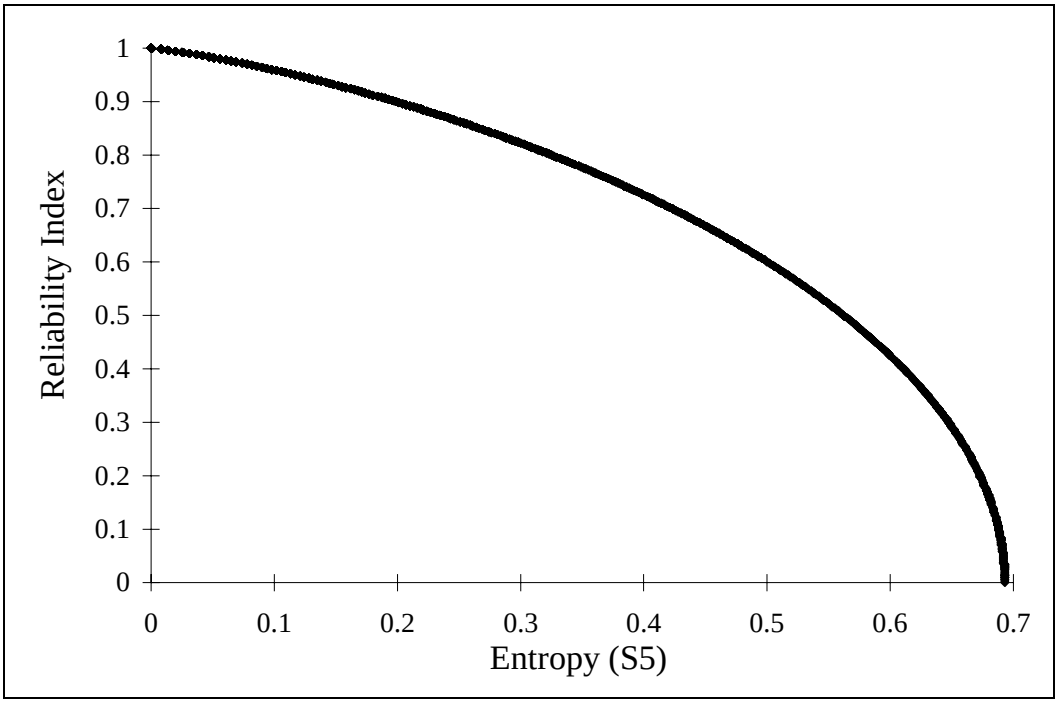

Fig.1. Relationship between the reliability index (defined as R=| O(H) - O(NH) |, where the Os are the network outputs for the helix and non-helix predictions (Rost and Sander, 1994)) and the Shannon entropy (Eq. 1) computed from the outputs for each predicted residue of the data set.

Evaluating the entropy threshold value

The predictive accuracy of the network, albeit good (for comparison see a somewhat similar predictor in Rost and Sander, 1993, characterized by an accuracy (Q2) of 81%) is unfortunately not perfect (Table 1). This is apparent from Fig. 2, where the distribution of the entropy values are shown both for the segments correctly predicted in the helix state and for those not correctly predicted (one fragment is not correctly predicted provided that it contains at least one residue with a wrong prediction). At low entropy values, correct predictions are more frequent than the wrong ones. The distributions of the correct and wrong predictions peak at different entropy values (S5=0.139 and S5=0.554, respectively); however a significant overlap of the distributions is noticeable. Maximization of the correct predictions over the wrong ones is obtained by introducing a threshold entropy value. This is evaluated by considering that at the intersection of the two distributions (Fig 2) the frequency of the correct predictions equals that of the wrong ones. The threshold value S5 is therefore set equal to the intersection entropy value (0.416). As a consequence, only those helical segments which are characterized by an entropy value ≤0.416 are accepted and included in the data base of minimally frustrated protein fragments. This S5 value corresponds to a Pc(H) value equal to 0.85 and a reliability index equal to 7 (Rost and Sander, 1994).

Comparison with experimental results

A typical smoothed entropy profile of the alpha helical predictions in a protein chain is shown in Fig.3. The protein analyzed is the hen egg white lysozyme whose putative initiation sites have been elucidated by NMR (Radford et al., 1992). The S5 minima correspond to the folding initiation sites experimentally found. Entropy minima which are not in the helix state are not considered by our procedure. The former analysis carried out in Compiani et al. (1998) is here extended to a new set of putative folding initiation sites of proteins (Table 2). Unfortunately the presently available data on experimentally detected initiation sites are still largely insufficient to determine a S5 cut-off threshold value for accepting minimally frustrated helical segments. Yet the large majority (>95%) of the about 50 putative initiation sites experimentally detected in native proteins is successfully predicted by our method.

In Table 3, the predictor is tested on a number of segments whic h are self-stabilizing, namely protein

fragments which are documented to have some helical content also in polar solution (Muñoz and Serrano, 1994). A good correlation (r=0.8) is noticeable between the fragments extracted by our predictor and the helical protein fragments experimentally detected.

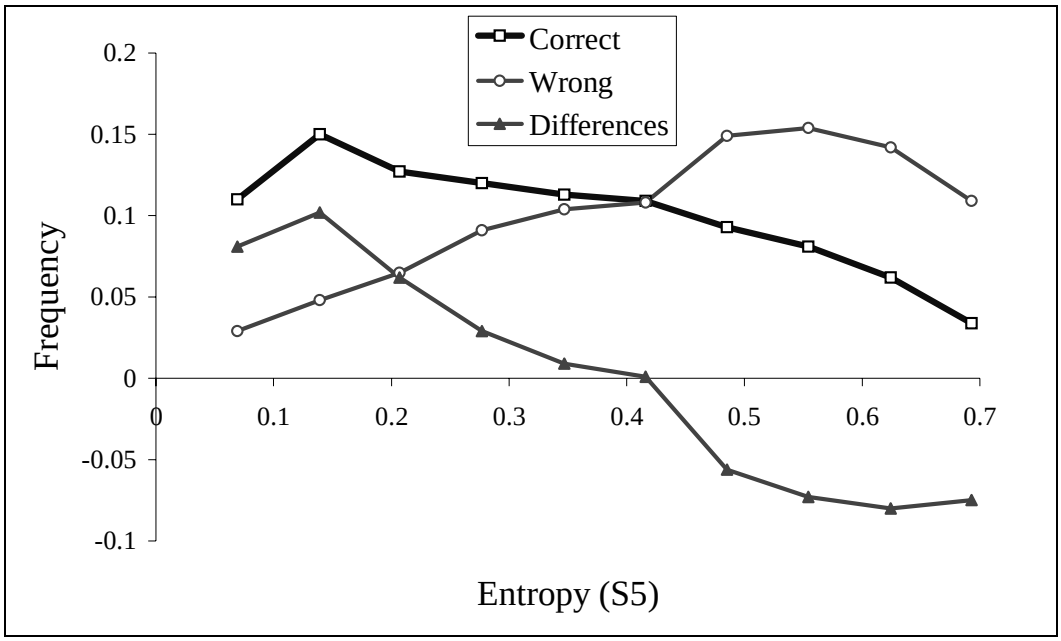

Fig.2. Frequency distribution of predicted helical segments as a function of their entropy value averaged on a segment of 5 contiguous residues (S5). The difference between the correct and wrong predictions highlights the intersection entropy value (0.416) which is used as a threshold value for defining minimally frustrated segments.

Table 2: Comparison of minimally frustrated segments with putative folding initiation sites experimentally determined.

PDB Code	Entropy (S5)	Position in the protein chain	Extracted Sequence	Reference
132l	0.109	8-14	LAAAMKR	Radford & al.(1992)
132l	0.212	89-95	TASVNCA	Radford & al.(1992)
1hfx	0.186	86-91	TDDIMC	Chyan & al. (1993)
1hfx	0.221	7-13	ALSHELN	*
1hrc	0.156	92-99	EDLIAYLK	Jeng & al. (1990)
2mm1	0.050	127-132	AQGAMN	Hughson & al. (1990)
2mm1	0.104	139-146	RKDMASNY	Hughson & al. (1990)
2mm1	0.154	105-111	EFISEAI	Hughson & al. (1990)
7rsa	0.409	8-11	FERQ	Udgaonkar & Baldwin(1990)
1ubq	0.322	25-28	NVKA	Briggs & Roder (1992)
1gf1	0.311	10-16	LVDALQF	Hua & al. (1996)
2ci2	0.236	14-19	VEEAKK	Fersht (1995)

*Not yet experimentally detected

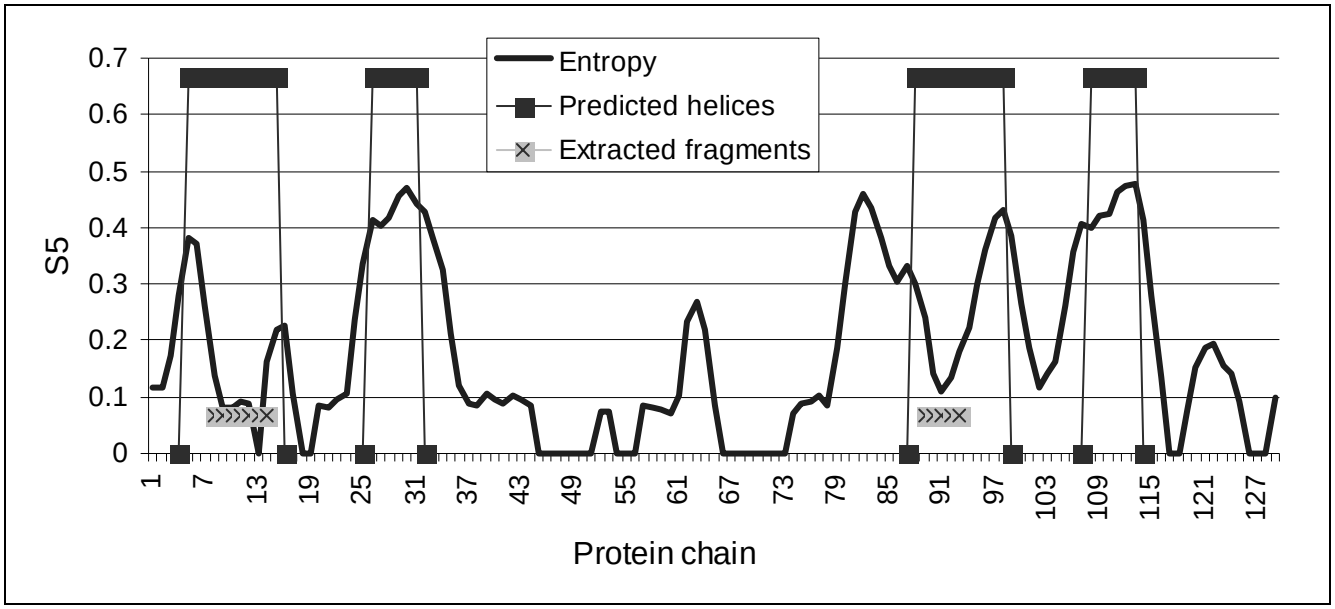

Fig.3 Profile of the smoothed entropy (S5) for the hen lysozyme (132L). Two entropy minima are extracted from the four correctly predicted helical segments (Q2=92%), and these have also been experimentally identified as putative protein folding initiation sites (Radford et al., 1992).

The Data Base of minimally frustrated segments

The structure of the data base is as simple as possible to facilitate the visualization and downloading of the data from our web site: http://www.biocomp.unibo.it. The extracted segments are organized in text format in two different tables, which share the same structure and are arranged in records. Records are sorted differently. Each record contains five fields:

- the protein PDB code, with the indication of the chain that has been analyzed;
- the entropy value of the extracted segment;
- the starting and ending positions in the protein chain of the extracted segment;
- the amino acidic residue sequence (in one letter code) of the extracted segment;
- the assigned structure according to the DSSP program.

An excerpt of the data base is shown in Fig.4. The protein fragments are sorted according to their entropy values starting from the lowest ones. The lower the entropy the higher the probability that the helical segments are minimally frustrated. The second type of table (not shown) lists the segments sorted by the PDB codes of the protein from which they have been extracted. In this case it is possible to search for segments related to a particular protein.

The main characteristics of the helical segments contained in the data base are summarized in the following. With our minimum entropy criterion, 626 proteins of the training set contain reliable alpha helical fragments (or minimally frustrated fragments). Their length varies from 5 to 25 residues (average length = (7±2) residues, to be compared with the average length of the helical segments in the training sets (11±6)). 3000 fragments are listed with S5≤ 0.416. Among these, only 478 (16% of the total) have at least one residue in the segment which is not correctly predicted (these residues are indicated with exclamation marks in the fragments as assigned by the DSSP program). Further information can be derived from the data base, such as the frequency of occurrence of the residues in the data base as compared to that in the helices of the training set and the frequency of occurrence of paired residues in position i and respectively i+2, i+3 and i+4. In Fig. 5, the frequency of the twenty different residues is plotted with respect to the relative abundance of the residues in the helices of the training set, starting from the most frequent residue (A) to the less frequent one (P). This distribution is compared to that of the residues in the protein helical fragments of the data

base. The frequency difference between the two distributions indicates that the residue composition of the data base amplifies the general trend of the residue composition of the helix training set. Considering the physico-chemical properties of the majority of residues characterized by positive values of the frequency difference, it can be concluded that on average the fragments of the data base are more hydrophobic than the helical segments of the training set. In Fig.6, the frequency of occurrence of the paired residues in position i and i+4 in the fragments of the data base is shown after normalization to the same value computed from the helices of the training set. It appears that the most emerging residue pairs are EE, IC, DH, LI and IL (shown in black in Fig.6). This indicates that the most reliable patterns predicted by the network with a low entropy value are helix blocks where local stabilization within one helix turn is determined by side chain interaction of not only hydrophobic residues (IL, LI), but also of polar and apolar residues (IC), charged and polar residues (DH) and charged residues (EE). Careful analysis of the 101 EE pairs extracted from 89 proteins of the data base shows that most of them (>96%) are exposed to the polar solvent in the native protein. Likewise 88% of the 17 DH pairs (from 16 proteins) are exposed to solvent in the folded protein. EE pairs (in positions i, i+4) are seemingly compatible with the high helix content (measured experimentally) of the CIIIL and COMA4 peptides in Table 3. These results, all together, provide useful hints for determining helix building blocks for the rational design of peptides with helical propensity.

Table 3: Comparison of minimally frustrated segments with peptides extracted from proteins

Code*	Peptides*	% Helix in solution*	Entropy (S5)	Extracted Segment
3FXC	TYKVTELINEAEGINETIDCDD	1	#####	####
3LZM	GFTNSLRMLQQKRWDEAVNLAKS	10	0.262	WDEAVNL
	"	10	0.329	LRMLQQK
3LZM-2	GVAGFTNSLRMLQQKRWDEAAVNLAKS	12	0.203	SLRMLQ
	"	12	0.210	DEAAVNL
CIII	ESLLERITRKLRDGWKRLIDIL	8	0.171	LLERIT
	"	8	0.260	WKRLID
CIII-L	ESLLERITRKL	15	0.171	LLERIT
CIII-R	RDGWKRLIDIL	4	0.260	WKRLID
CIII-M	RITRKLRDGWK	2	####	####
Sigma	KVATTKAQRKLFFNLRKTKQRL	9	0.218	TKAQRK
COMA1	DHPAVMEGTKTILETDSNLS	4	####	####
COMA2	EPSEQFIKQHDFSSY	3	####	####
COMA3	VNGMELSKQILQENPH	6	0.189	LSKQILQ
COMA4	EVEDYFEEAIRAGLH	20	0.020	YFEEAIR
COMA5	KEKITQYIYHVLNGEIL	3	####	####
ARA1	AVGKSNLLSRYARNEFSA	2	####	####
ARA2	RFRAVTSAYYRGAVG	3	####	####
ARA3	TRRTTFESVGRWLDELKIHSD	7.5	0.194	SVGRWL
ARA4	AVSVEEGKALAEEEGLF	4	####	####
ARA5	STNVKTAFEMVILDIYNNV	3	####	####
G1	DTYKLILNGKTLKGETTTEA	2	####	####
G2	GDAATAEKVFKKIANDNGVD	4	####	####
G3	GEWTYDDATKTFTVTE	2	####	####

* Protein fragments whose alpha helical content in polar solution was determined by means of circular dicroism (Muñoz and Serrano, 1994). Similar values of alpha helical content are predicted by the AGADIR algorithm (http://www.embl-heidelberg.de/Services/Serrano/agadir/agadir-start.html) estimating the helical behavior of monomeric peptides in solution. Extracted segment = the segment extracted by the predictor. Entropy= value of the minimal entropy averaged on 5 neighboring residues. ####= no pattern within the entropy threshold is extracted by the predictor.

CODE	ENTROPY	POSITIONS	SEQUENCE	DSSP SECONDARY STRUCTURE
1msk_	0.002	192-206	ADRLAEAFAEYLHER	HHHHHHHHHHHHHHH
1pyda	0.004	307-319	MKFVLQKLLTNIA	HHHHHHHHHHHHH
1ngr_	0.005	63-72	LDALLAALRR	HHHHHHHHHH
1sly_	0.005	338-346	AKEILHQLM	HHHHHHHH!
1aerb	0.006	20-28	VERLLQAHR	HHHHHHHHH
1bcn_	0.006	113-123	LENFLERLKTI	HHHHHHHHHHH
1bib_	0.006	215-226	LAAMLIRELRAA	HHHHHHHHHHHH
1fkx_	0.006	337-346	KKELLERLYR	HHHHHHHHHH
2arcb	0.006	148-158	NLLEQLLLRRM	HHHHHHHHHHH
1aqt_	0.008	112-125	DYAQASAELAKAIA	HHHHHHHHHHHHHH
1fit_	0.008	111-120	EEEXAAEAAA	HHHHHHHHHH
1mtyg	0.009	22-30	LEKAAEMLK	HHHHHHHHH
2tct_	0.009	50-60	LLDALAVEILA	HHHHHHHHHHH
1hsba	0.010	150-157	AHVAEQWR	!!HHHHHH
2chsa	0.010	17-26	EEILQKTKQL	HHHHHHHHHH
1hjp_	0.011	175-184	ETLIREALRA	HHHHHHHHH!
1pou_	0.011	5-13	LEQFAKTFK	HHHHHHHHH

.....

Fig. 4: An example of the data base of minimally frustrated protein fragments sorted by their minimum entropy value. Non helical residues in the corresponding protein chain (as defined by the DSSP program) are indicated with an exclamation mark and correspond to wrong predictions given by the network.

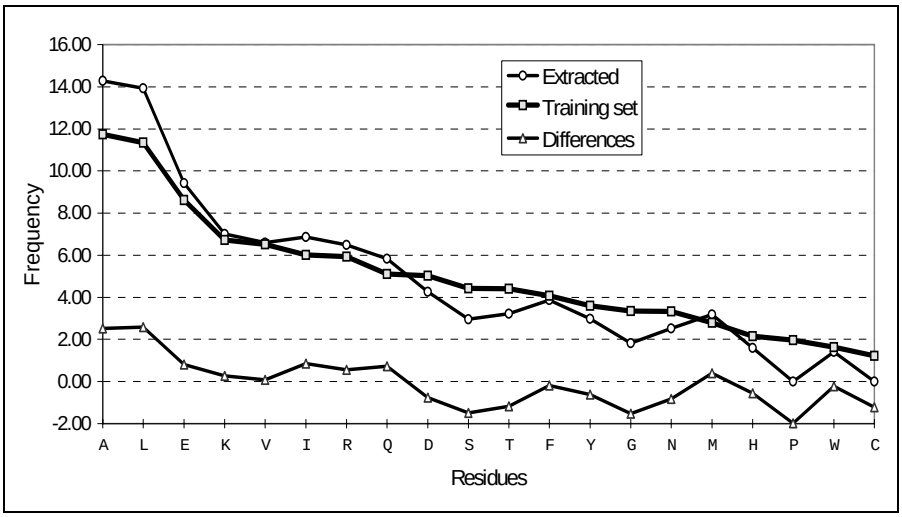

Fig.5. The frequency distribution of the 20 amino acid residues in the data base of minimally frustrated segments as compared to that of the α-helix structures in the training set. The difference curve highlights those residues which are more frequent in the data base as compared to the training set.

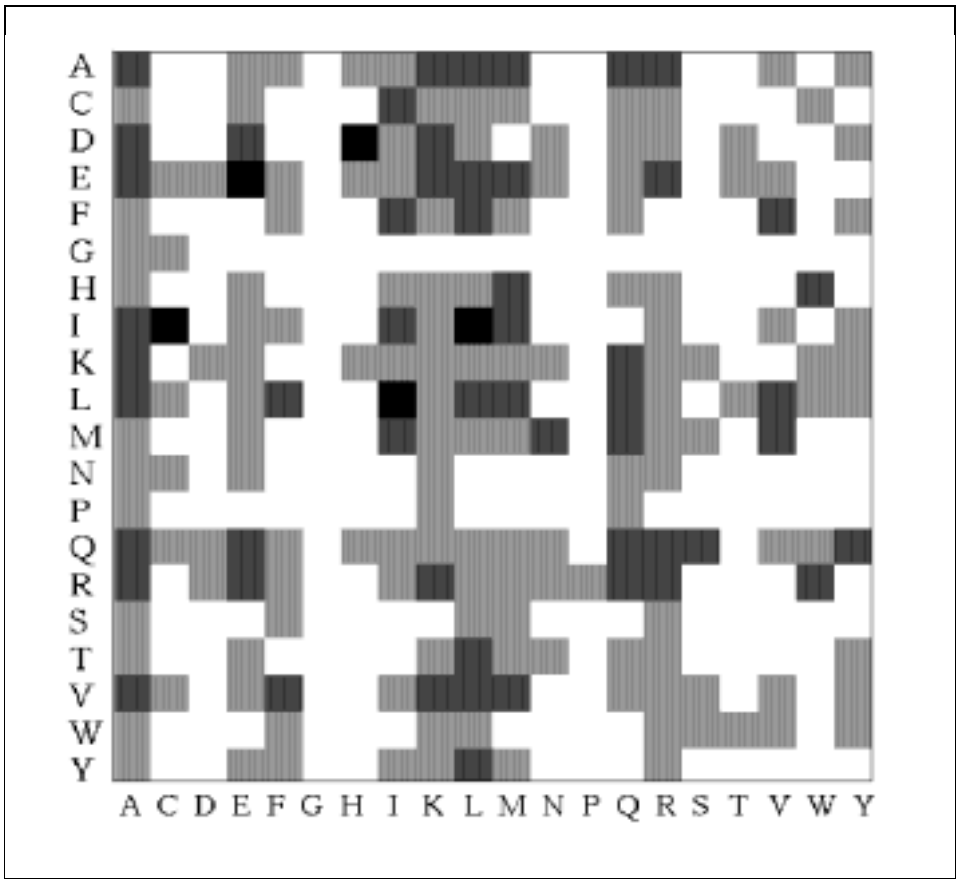

Fig.6 . Frequency distribution of paired residues in position i (y-axis) and i+4 (x-axis) in the helical fragments of the data base normalized to the analogous distribution of the α-helix structures in the training set. Black: ratio > 1.5; dark gray: ratio ≤ 1.5, ≥ 1.2; light gray: ratio < 1.2, ≥ 0.8; white: ratio < 0.8.

Remarks

For proteins not belonging to our training set, the predictor is available upon request.

Acknowledgments

Financial support to this work was provided by a grant of the Ministero della Pubblica Istruzione e della Ricerca Scientifica e Tecnologica (MURST) delivered to the project “Biocatalisi e Bioconversioni” and by a grant for a target project in Biotechnology of the Italian Centro Nazionale delle Ricerche (CNR).

References

Abkevich, V.I.; Gutin, A.M. and Shakhnovich, E. 1994. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* 33:10026-10036.

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

Bishop, C.M. 1994. Neural networks and their applications. *Rev. Sci. Instrum*. 65:1803-1832.

Briggs, M.S.; and Roder, H. 1992. Early hydrogen bonding events in the folding reaction of ubiquitin. *Proc. Natl. Acad. Sci U S A* 89:2017-2021.

Bryngelson, J.D.; and Wolynes, P.G. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U S A* 84:7524-7528.

Bryngelson, J.D.; Onuchic, J.N.; Socci, N.D.; and Wolynes P.G. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21:167-95.

Chyan, C.-L.; Wormald, C.; Dobson, C.M.; Evans, S.W.; and Baum, J. 1993. Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: a hydrogen exchange study. *Biochemistry* 32:5681-5691.

Compiani, M.; Fariselli, P.; Martelli, P.; and Casadio, R. 1998. An entropy criterion to detect minimally frustrated intermediates in native proteins. *Proc. Natl. Acad. Sci.* USA 95:9290-9294.

Cuff J.A.; and Barton G.J. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 34:508-519.

Dill, K.A.; and Chan, H.S. 1997. From Levinthal to pathways to funnels. *Nature Struct. Biol.* 4:10-19.

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.

Frauenfelder, H.; and Wolynes P.G. 1994. Biomolecules: where the physics of complexity and simplicity meet. *Physics Today* 47:58-64.

Hua, Q.X.; Narhi, L.; Jia, W.; Arakawa, T.; Rosenfeld, R.; Hawkins, N; Mileer, J.A.; and Weiss, M.A. 1996. Native and non native structures in a protein folding intermediate; spectroscopic studies of a partially reduced IGF-I and an engineered alanine model. *J. Mol. Biol.* 259:297-313.

Hao, M.-H. ; and Scheraga, H.A. 1998. Molecular mechanisms for cooperative folding of proteins. *J. Mol. Biol.* 277:973-983.

Hughson, F.M.; Wright, P.E.; and Baldwin, R.L.1990. Structural characterization of a partly folded apomyoglobin intermediate. *Science* 249:1544-1548.

Jeng, M.F.; Englander, S.W.; Elove, G.A.; Wand, A.J.; and Roder, H. 1990. Structural description of acid denaturated cytochrome c by hydrogen exchange and 2D NMR. *Biochemistry* 29:10433-10437.

Kabsch, W.; and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.

Karplus, M.; and Weaver, D.L. 1994. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.* 3:650-668.

Klimov, D.K.; and Thirumalai, D. 1998. Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol*. 282:471-492.

Mamitsuka H.; and Yamanishi K. 1995. α-helix region prediction with stochastic rule learning. *CABIOS* 11:399-411.

Matthews, C.R. 1993. Pathways of protein folding. *Ann. Rev. Biochem.* 62:653-683.

Moult, J.; and Unger, R. 1991. An analysis of protein folding pathways. *Biochemistry* 23:3816-3824.

Muñoz, V.; and Serrano, L. 1994. Elucidating the folding problem of helical peptides using empirical parameters. *Nature Struct. Biol.* 1: 399-409.

Muñoz, V.; and Serrano, L. 1995. Helix design, prediction and stability. *Curr. Op. Biotech.* 6: 382-386.

Presta, L.G.; and Rose, G.D. 1988. Helix signals in proteins. *Science* 240:1632-1641.

Qian, H.; and Chan, S.I. 1996. Interactions between a helical residue and tertiary structures: helix propensities in small peptides and in native proteins. *J. Mol. Biol.* 261:279-288.

Radford, S.E.; Dobson, C.M.; and Evans, P.A. 1992. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* 358:302-307.

Rooman, M.J.; Kocher, J-P.A.; and 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.

Rost, B.; and Sander, C. 1993. Secondary structure prediction of all-helical proteins in two states. *Prot. Engng*. 6:831-836.

Rost, B.; and Sander, C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19:55-72.

Rumelhart, D.E.; Hinton, G.E.; and Williams, R.J. 1986. Learning representations by back-propagating errors. *Nature* 323:533-536.

Sali, A.; Shakhnovich, E.; and Karplus, M. 1994. How does a protein fold? *Nature* 369:248-251.

Sander, C.; and Schneider, R. 1991. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 9:56-68.

Scholtz, J.M.; and Baldwin, R.L. 1992. The mechanism of α-helix formation by peptide. *Annu. Rev. Biophys. Biomol. Struct.* 21:95-118.

Serrano, L.; Matouschek, A.R.; and Fersht, A.R. 1992. The folding of an enzyme VI: the folding pathway of barnase. Comparison with theoretical models. *J. Mol. Biol.* 224:847-859.

Udgaonkar, J.B.; and Baldwin, R.L. 1990. Early folding intermediates in ribonuclease A. *Proc. Natl. Acad. Sci. U S A.* 87:8197-8201.

Zhang, C.T.; and Chou, K.C. 1992. An optimization approach to predicting protein structural class from amino acid composition. *Prot. Sci.*1:401-408.