Predictions of Protein Segments With the Same Aminoacid Sequence and Different Secondary Structure: A Benchmark for Predictive Methods

Irene Jacoboni,[1] **Pier Luigi Martelli,**[1] **Piero Fariselli,**[1,2] **Mario Compiani,**[1] **and Rita Casadio**[1,2*]

[1]*Laboratory of Biocomputing, Centro Interdipartimentale per le Ricerche Biotecnologiche (CIRB), Bologna, Italy*

[2]*Laboratory of Biophysics, Department of Biology, University of Bologna, Bologna, Italy*

ABSTRACT **The most stringent test for predictive methods of protein secondary structure is whether identical short sequences that are known to be present with different conformations in different proteins known at atomic resolution can be correctly discriminated. In this study, we show that the prediction efficiency of this type of segments in unrelated proteins reaches an average accuracy per residue ranging from about 72 to 75% (depending on the alignment method used to generate the input sequence profile) only when methods of the third generation are used. A comparison of different methods based on segment statistics (2nd generation methods) and/or including also evolutionary information (3rd generation methods) indicate that the discrimination of the different conformations of identical segments is dependent on the method used for the prediction. Accuracy is similar when methods similarly performing on the secondary structure prediction are tested. When evolutionary information is taken into account as compared to single sequence input, the number of correctly discriminated pairs is increased twofold. The results also highlight the predictive capability of neural networks for identical segments whose conformation differs in different proteins. Proteins 2000;41:535–544.** **© 2000 Wiley-Liss, Inc.**

Key words: neural networks; secondary structure predictions; multiple sequence alignment; pattern recognition; chameleon sequences

INTRODUCTION

Theoretical and experimental studies of protein folding indicate that the protein native structure involves a delicate balance between local and non-local interactions.[1,2] Local interactions lead to the formation of stable fragments of secondary structures, rather independently of the protein global context, whereas non-local factors are responsible for the overall formation of the stable tertiary structure.[3–7] This view is corroborated by the finding that the database of proteins known with atomic resolutions contains unrelated chains with short identical sub- sequences that are endowed with different secondary structure depending on the protein global context.[8–12] Segments with identical sequence and different conformations are referred to as “chameleon” sub-sequences.[4]

The search of different releases of the ever-increasing database of protein structures (PDB)[13] has shown that identical k-mers of different length (with $5 \leq k \leq 8$) can be found with different conformations in couples of unrelated proteins (whose sequence identity is <25%).[8–12] The presence of these segments blurs the structural assignment of a given residue and corrupts the pattern classification into structural classes.[14] It can be regarded, therefore, as a limiting factor of the predictive performance of all the computational methods that aim at the prediction of secondary structure starting from the sequence. These methods include the residue local context using a sliding input window to perform segment statistics during the training phase[14–18] (they are referred to as the 2nd generation methods[19]). It has been proven that methods taking into account evolutionary information (3rd generation methods[19]) can largely increase the predictive performance of secondary structure prediction.[18,20] Systems based on a consensus procedure can perform even better when a systematic comparison of all the top scoring methods is performed.[21] One interesting question that can be posed at this stage of the development of secondary structure prediction is to what extent predictors are capable of distinguishing the different structures of chameleon sequences. This will provide insights into the predictive performance and will highlight to what extent chameleon sequences are endowed with a wrong prediction. Evidently, these segments are blurring the mapping computed by the networks to a larger extent than those patterns endowed with a unique sequence-to-structure relation.[14]

Presently, 3rd generation predictive methods are characterized by an accuracy ranging from 72–78% (see Jones[25]) depending on the type of predictor, alignment method, and structural mapping.[21–25] The task of predicting chameleon sequences can, therefore, be posed as a benchmark to test the discriminative capability of these top-scoring predic-

Contract grant sponsor: Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST); Contract grant sponsor: Italian Centro Nazionale delle Ricerche (CNR).

*Correspondence to: Rita Casadio, Department of Biology, University of Bologna, Via Irnerio 42, 40126 Bologna, Italy. E-mail: casadio@kaiser.alma.unibo.it;

Received 25 April 2000; Accepted 4 August 2000

tive methods. Several Web sites are presently available for secondary structure prediction. The performance of the methods, as described in the literature, is quite accurate.[21–26] However, since it is not always possible to trace the protein sets used for the training phase of the different methods, we implemented our predictor based, as the majority of the well-performing predictors, on neural networks. This allowed us to perform a reliable cross-validation discriminating between training and testing sets during our study.

We searched a subset of the PDB database containing a large number of protein chains with low identity (<25%) for those chameleon sequences with a variable length ranging from 5 to 8 residues and totally different conformations. The prediction of these segments was then extracted from the prediction of the correspondent proteins predicted with our method and other top-scoring ones. Our results indicate that 3rd generation methods are superior in predicting chameleon sequences as compared to 2nd generation methods. Noticeably, it appears that 2nd generation methods predict chameleon sequences with an accuracy that is lower than the predictive accuracy of the method, indicating that the predictor is much affected by the ambiguity that blurring patterns introduce in the sequence-to-structure mapping. When evolutionary information is taken into account chameleon sequences are predicted with efficiency similar to that of the predictive method on the global testing set. This clearly indicates that sequence profile is sufficient to partially compensate for blurring. The results are rather independent of the predictive method and seem to be somewhat affected by the procedure used to generate multiple alignment.

MATERIALS AND METHODS

Protein Database

The neural network based predictor is trained on protein chains with a low level of identity (<25%). To avoid redundancy in the training-set, a data set of 822 proteins known at a molecular level (and containing 174,192 residues) is derived from the database of non-homologous proteins (with an identity value <25%) using the PDB_select_jun_98 algorithm (http://www.embl-heidelberg.de). The Swiss-Prot database of known protein sequences (Release 38, July 1999) is used for pairwise and multiple alignment of each of the query sequences.[27] Secondary structure assignment is done with the DSSP.[28] This program defines 8 states for secondary structure (H, E, B, T, S, L, G, and I)[28] that are reduced to three states, H, E, and C, by different predictive methods.[21–26] In assigning secondary structure, we used the following reduction: H and G to helix (H), E and B to beta strands (E), all the rest to coil (C). It is very well documented that this three-state reduction affects the predictor accuracy and that it promotes a lower accuracy than classifying G in C.[21]

Database of Chameleon Sequences

A program written in C language is implemented to search the selected database for segments with identical sequence and different secondary structure. Secondary structure is assumed to be different if no amino acid in the segments has the same secondary structure in the same position (such as the pair HHHHHH and EEEEEE, or the pair CCCHHH and HHHEEE).

The former procedure selects 2,452 couples of segments comprising 5 residues (5-mers), 107 couples comprising 6 residues (6-mers), and 12 couples including 7 residues per segment (7-mers). Other couples, found in the literature, that satisfy our criterion for structural diversity are also added to the database of chameleon sequences, and predicted, when necessary, by our predictor with cross-validation: 1bgw-1mdaH; 1cgu-1bglA; 1thg-1igmH taken from Argos[9]; 1pgs-2sblB and 1pht-1wbc taken from Cohen et al.[10]. In this way, the complete database includes 2,452 couples of 5-mers, 107 couples of 6-mers, 16 couples of 7-mers, and one only couple containing 8 residues (8-mers), in sum 2,576 couples, a set much larger than those previously reported.[8–12] An accurate search of the database and of previously reported data on chameleon sequences did not increase the number of couples to be included in the 8-residue-long category. The total number of residues is equal to 26,044, out of 755 proteins. The sequences, structures, PDB identification codes, and solvent accessibility values of the 6-, 7-, and 8-mer couples are listed below.

Solvent Accessibility

The solvent accessibility of each segment is the solvent accessibility value per residue as computed by the DSSP program[28] averaged over the segment length. Solvent accessibility per residue is evaluated by normalizing the computed value to the maximal exposed surface area of the residue[29] in the database of selected proteins. Two categories of segments (buried and exposed) are discriminated depending on the average accessibility value being higher and lower (or equal to) than a 16% threshold.[30] This is a limiting discriminating value for classifying a residue buried (<16%) or exposed (≥16%).[30]

Neural Network-Based Predictor

A feed-forward neural network is implemented and trained with the back propagation algorithm.[31] The network architecture basically consists of perceptrons with one hidden layer containing 22 hidden nodes and an input window spanning 17 residues. Three output nodes are considered in order to discriminate three structural types: alpha (H), beta (E), and coil motifs (C) of secondary structure. The architecture of the predictor is extended to include a second cascaded network to filter out spurious assignments (a so-called structure-to-structure step[20]).

The prediction is finally obtained by averaging over the outputs of six different predictors (all based on the architecture described above) acting as a jury. The six predictors include (1) different window lengths (9 and 17 residues); (2) weight balancing during training[32]; (3) distinguishing two structural types instead of three. Each of the predictors was trained with a 20-fold cross-validation on the 822 proteins selected from the PDB.

Evolutionary information is given as input in the form of sequence profiles after multiple sequence alignments.

TABLE I. Efficiency of the Neural Network-Based Predictors on the 822 Proteins of the Testing Set

Input[a]						
Single Sequence			Q3 (%)	66.3		
			SOV	0.62		
	Q[H]	0.69	Q[E]	0.61	Q[C]	0.66
	P[H]	0.70	P[E]	0.54	P[C]	0.71
	C[H]	0.54	C[E]	0.44	C[C]	0.45
Multiple sequence (MaxHom)			Q3 (%)	72.4		
			SOV	0.69		
	Q[H]	0.75	Q[E]	0.65	Q[C]	0.75
	P[H]	0.77	P[E]	0.64	P[C]	0.73
	C[H]	0.64	C[E]	0.54	C[C]	0.53
Multiple Sequence (PSI-BLAST)			Q3 (%)	73.4		
			SOV	0.70		
	Q[H]	0.75	Q[E]	0.70	Q[C]	0.73
	P[H]	0.80	P[E]	0.63	P[C]	0.75
	C[H]	0.67	C[E]	0.56	C[C]	0.53

[a] Input to the networks included single sequence or sequence profiles evaluated with MaxHom[33] and PSI-BLAST,[34] respectively. Scores are computed with a cross-validation procedure. The different statistical indexes are defined in the Appendix.

Sequence alignments were derived from the HSSP database[33] in which alignments were constructed using BLAST[34] to search the sequence database and MAXHOM[35] to align the sequences. Moreover we used PSI-BLAST[36] (3 rounds with threshold equal to 0.001) to search the Swiss-Prot database and we generated sequence profiles from its outputs by means of a newly implemented program. This is based on the notion that the PSI-BLAST complete outputs contain the local pairwise alignments of the query sequence with all the extracted sequences. From this it is possible to compute a profile by merging each local pairwise alignment.[37] This second alignment method, as compared with the first, gave a 1% increase of the overall efficiency on the 822 proteins. The overall performance of our predictor (available at www.biocomp.it) is shown in Table I, both using single and multiple sequence as input. The overall efficiency of this predictor using multiple sequence is somewhat lower than that of another recently published method also based on the use of PSI-BLAST and neural networks.[25] In our opinion, this is possibly due to the different databank used for homology search (Swiss-Prot containing about 80,000 sequences in this study against a selected databank of 340,000 chains mentioned in Jones[25]). In Figure 1 the distribution of the accuracy per protein obtained using as input single and multiple sequence is reported for the sake of comparison with previous work.[25] Prediction efficiency is evaluated by computing different scoring indexes (see Appendix[38]).

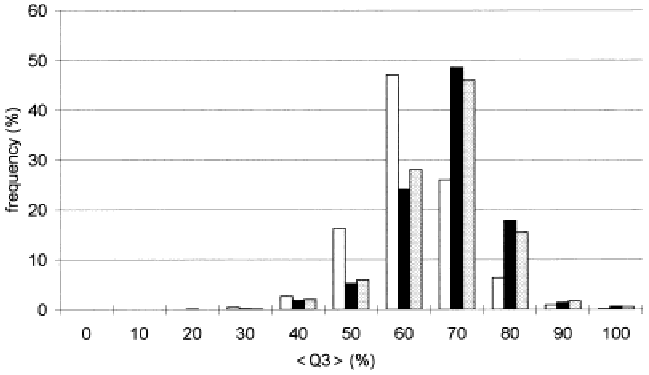

Fig. 1. Distribution of Q3 scores for the 822 proteins predicted by the neural networks in cross-validation. Distribution of accuracy is shown for predictions computed using single sequence (white bars) and sequence profiles (compared with MaxHom, grey bars) and PSI-BLAST (black bars) as input to the networks.

RESULTS AND DISCUSSION

Characterizing Chameleon Sequences

The residue composition of chameleon sequences is somewhat different from that of the protein database from where they have been extracted (Table II). Indeed, the relative frequency of occurrence of apolar residues is slightly higher than that of charged and polar ones. These results are in agreement with previous observations when chameleons were extracted with other methods from databases of proteins much smaller than ours.[11] They also seem to corroborate the suggestion that alanine, valine, isoleucine, and leucine taken in any pair have the most chance to produce favourable interactions under a variety of different circumstances.[39] Moreover, cysteine, triptophan, methionine, proline, and hystidine residues are significantly less abundant in chameleon segments as compared to the protein sequences. It appears that in our relatively large database of chameleons, residues are non-uniformly distributed and this suggests that the structural adaptability of proteins should vary from sequence to sequence.[8]

When secondary structures are determined with DSSP, about 39% of the couples are found either in alpha helical motifs (H) or in mixed coil-strand (C-E) structures, 16%

TABLE II. Frequency of Occurrence of Amino Acids in Chameleons and in the Protein Database[a]

	G	A	V	F	P	M	I	L	S	T	Y	H	C	N	Q	W	D	E	K	R
f ch	7.93	12.67	10.51	2.98	2.20	0.94	6.42	13.33	5.24	5.61	2.28	1.18	0.46	2.92	2.70	0.46	5.07	6.87	6.04	4.20
f pr	7.62	8.09	6.93	4.07	4.79	2.17	5.54	8.38	5.96	5.78	3.69	2.33	1.64	4.61	3.82	1.55	5.95	6.24	5.94	4.86
f	1.04	1.57	1.52	0.73	0.46	0.43	1.16	1.59	0.88	0.97	0.62	0.51	0.28	0.63	0.71	0.30	0.85	1.10	1.02	0.86
	Apolar								Polar								Charged			

[a] f ch: frequency of occurrence (%) of amino acids in chameleons; f pr: frequency (%) of occurrence of amino acids in the database of 822 proteins; f: f ch/f pr. Based on multinomial distribution, f values >1.1 or <0.9 are statistically significant.

TABLE III. Motifs of Secondary Structure and Solvent Accessibility of Chameleon Couples

A. 5-mers

Segment 1 struct. type	Segment 2 struct. type	No[a]	No (both exposed)[a]	No (mixed)[a]	No (both buried)[b]
C	E	140	83	45	12
C	H	390	294	83	13
E	H	340	150	142	48
C	E-H	21	14	7	0
E	C-H	240	130	92	18
H	C-E	952	583	287	82
C-E	C-H	263	179	71	13
C-E	C-E	22	11	10	1
C-E	E-H	17	11	5	1
C-H	E-H	6	4	2	0
C-H	C-H	30	26	4	0
C-E	C-E-H	11	5	5	1
C-H	C-E-H	17	13	4	0
E-H	C-E-H	2	2	0	0
C-E-H	C-E-H	1	1	0	0

B. 6-, 7-, 8-mers

Segment 1 struct. type	Segment 2 struct. type	No[b]	No (both exposed)[b]	No (mixed)[b]	No (both buried)[b]
C	E	4	2	1	1
C	H	17	13	3	1
E	H	10	3	6	1
C	E-H	2	2	0	0
E	C-H	13	4	8	1
H	C-E	59	33	21	5
C-E	C-H	16	7	8	1
C-E	E-H	2	2	0	0
C-H	C-E-H	1	0	0	1

[a] Number of chameleon couples. The total number of 5-mer couples is 2,452.
[b] Number of chameleon couples. The total number of 6-, 7-, 8-mer couples is 124.

are either in H or in C, and 14% are either in H or in E structures. The remainder is distributed over the other possible structures (Table III). Therefore, chameleons seem to predominantly adopt H/C-E secondary structures also when we consider the subset including only the 6-, 7-, and 8- mers (for a total of 124 couples) (Table III).

We can ask the question as to where chameleons are located in the protein with respect to the solvent. The analysis of solvent accessibility indicates that a majority of the couples are composed of exposed segments (61%), that 31% of the couples are mixed with one segment exposed and the other buried, and that only 8% are buried (Table III). This trend is rather independent of the structural compositions of the couples. In conclusion, our results show that in a large database of non-redundant proteins, chameleon sequences (comprising about 15% of the whole set of residues) exhibit general structural features that are also shared by the remainder of the protein residues.

Predicting Chameleon Sequences

In order to predict a chameleon pair, first the two proteins to which the segments belong are predicted; then the prediction of chameleon sequences is extracted from that of the corresponding proteins. This is done by cross-validation both using single and multiple sequences as input to the networks. The prediction accuracy is evaluated using the scoring indexes listed in the Appendix (see Table IV). It is evident that prediction is significantly improved when evolutionary information is taken into

TABLE IV. Prediction of Chameleons With Neural Networks

5-mers						
Single sequence	Q3 (%)	58.6				
	Q[H]	0.69	Q[E]	0.55	Q[C]	0.47
	P[H]	0.65	P[E]	0.55	P[C]	0.51
	C[H]	0.42	C[E]	0.37	C[C]	0.29
Multiple sequence (MaxHom)	Q3 (%)	69.1				
	Q[H]	0.78	Q[E]	0.63	Q[C]	0.63
	P[H]	0.78	P[E]	0.68	P[C]	0.58
	C[H]	0.62	C[E]	0.52	C[C]	0.43
Multiple sequence (PSI-BLAST)	Q3 (%)	71.3				
	Q[H]	0.80	Q[E]	0.69	Q[C]	0.61
	P[H]	0.81	P[E]	0.69	P[C]	0.60
	C[H]	0.66	C[E]	0.56	C[C]	0.45
6-, 7-, 8-mers						
Single sequence	Q3 (%)	58.7				
	Q[H]	0.67	Q[E]	0.59	Q[C]	0.45
	P[H]	0.66	P[E]	0.55	P[C]	0.51
	C[H]	0.40	C[E]	0.39	C[C]	0.30
Multiple sequence (MaxHom)	Q3 (%)	71.6				
	Q[H]	0.79	Q[E]	0.71	Q[C]	0.60
	P[H]	0.82	P[E]	0.71	P[C]	0.57
	C[H]	0.66	C[E]	0.60	C[C]	0.42
Multiple sequence (PSI-BLAST)	Q3 (%)	75.1				
	Q[H]	0.83	Q[E]	0.79	Q[C]	0.60
	P[H]	0.86	P[E]	0.70	P[C]	0.64
	C[H]	0.72	C[E]	0.63	C[C]	0.48

account, particularly when PSI-BLAST is used to derive the protein profile (Table IV).

Using single sequence as input to the network, the prediction score per residue of chameleons is 7.6 percentage points lower than that obtained on the whole testing set with the same input procedure (see Table I). On the other hand, when sequence profiles are fed as input to the networks, prediction ranks 12 to 13 percentage points higher than using single sequence. In this respect, it is similar to the average prediction values obtained for the overall efficiency of the networks (72.4 and 73.4%, respectively, depending on the method used to generate sequence profile). Chameleons are, therefore, predicted with efficiency rather close to that of the method on the whole database. Ultimately, this clearly indicates that in spite of their intrinsic ambiguity chameleons are predicted with efficiency similar to that of the other protein segments. This is due to the length of the input window (17 residues long), which is apparently sufficient to compensate for the ambiguity of the chameleon subpatterns.

Data relative to the prediction and location of 6-, 7-, and 8- mer couples are shown separately in Table IV and also listed in Table V with the average solvent accessibility of each segment.

In Table VI, we used the same set of 6-, 7-, and 8- mers listed in Table V to test different methods presently available on the Web. We use the accuracy values obtained in training and testing by our predictor to settle the lower and higher limits of the performance both when single sequence and sequence profiles are used for prediction. If the method tested is performing better than ours, the expected accuracy value should be at least ≥58.9% in single sequence and ≥75% in multiple sequence (Table IV) whether the predictor includes the protein in the training set or not.

The results listed in Table VI point to several conclusions. It is evident that a neural network using single sequence as input is performing slightly better than a similar method (GOR IV)[40] based on the information theory and using all possible pair frequencies within a window of the same length as that of the neural network (17 residues long).

When evolutionary information is used, it appears that other methods based on neural networks perform similarly to our predictor (PHD,[18] PSI-PRED,[25] PRED2ARY[24]). A consensus-based method (JPRED[21]) also reaches a similar accuracy by means of a filtering procedure of methods that rank slightly worse when considered independently (DSC,[41] NNSSP,[23] PREDATOR,[22] as implemented in JPRED[21]).

Focusing on the Predictions of Chameleon Sequences

The effect of multiple sequence input on the structural discriminating capability of the networks is shown in Figure 2. The frequencies of occurrence of the average accuracy (<Q3>) obtained for the whole set of chameleon sequences is reported. Provided that evolutionary information is used, it is evident that if we allow at most one wrong prediction over the couple (<Q3> ≥ 90%), some 34% of the segments present in the database are correctly discriminated. In single sequence, this figure reduces to 15%. The result is particularly relevant if it is considered that we are

TABLE V. List of Secondary Structures, Predictions, and Sequences of Chameleons of Length Six, Seven, and Eight

Sequence	PDB id	Start res[a]	Structure	Prediction	Solv acc[b]	PDB id	Start res[a]	Structure	Prediction	Solv acc[b]	⟨*Q*3⟩ (%)
6-mers											
TVLETL	1h9_	11	EEEEEC	EEEEEC	E	1broa	82	HHHHHH	HHHHHH	E	100
RVPALV	1ak5_	125	HHHHHH	HHHHHH	B	3cox_	7	EEEEEE	EEEEEE	B	100
VDLLKN	1ao7b	38	EEEEEC	EEEEEC	E	1hjga	16	HHHHHH	HHHHHH	E	100
RYIELV	1atla	4	EEEEEE	EEEEEE	B	1opy_	14	HHHHHH	HHHHHH	E	100
AGVKKV	1bcmb	269	CCCCEE	CCCCEE	E	1glya	314	HHHHHH	HHHHHH	E	100
QLIIED	1bvh_	124	CCCCCC	CCCCCC	E	1ecmb	73	HHHHHH	HHHHHH	E	100
ASGQSY	1cem_	236	CCCCCC	CCCCCC	E	1nox	127	HHHHHH	HHHHHH	E	100
LLLQVA	1cnt2	69	HHHHHH	HHHHHH	B	2pola	33	EEEEEE	EEEEEE	E	100
EKVANL	1dkgb	5	HHHHHH	HHHHHH	E	1ptq_	44	CCCCCC	CCCCCC	E	100
NSILQR	1eur_	51	CEEEEE	CEEEEE	B	1xyza	31	HHHHHH	HHHHHH	E	100
SLLDEE	1garb	110	CCCCCC	CCCCCC	E	1pyta	79	HHHHHH	HHHHHH	E	100
VVNTMR	1gpb_	219	CEEEEE	CEEEEE	B	1kid_	80	HHHHHC	HHHHHC	E	100
TDVFIR	1hlb_	36	HHHHHH	HHHHHH	E	1lam_	198	EEEEEE	EEEEEE	E	100
AKLVAV	1ipwb	103	EEEEEE	EEEEEE	B	3sdha	135	HHHHHH	HHHHHH	E	100
VVTIEG	1ksr_	44	EEEEEC	EEEEEC	E	1ppn_	31	HHHHHH	HHHHHH	B	100
KVYIEK	1pyp_	16	EEEEEC	EEEEEC	E	5csma	119	HHHHHH	HHHHHH	E	100
GLRVLD	1rhs_	27	CEEEEE	CEEEEE	B	2dkb_	312	HHHHHH	HHHHHH	B	100
DTIALV	1rusa	194	HHHHHH	HHHHHH	E	2dri_	2	CEEEEE	CEEEEE	B	100
ITTVLN	1tys_	112	HHHHHH	HHHHHH	E	2sil_	200	CCCCEE	CCCCEE	E	100
VDLSHF	1uby_	178	CCCCCC	CCCCCC	E	2nef_	30	HHHHHH	HHHHHH	E	100
GKMVVT	1ytba	59	CEEEEE	CEEEEE	E	2rslc	104	HHHHHH	HHHHHH	E	100
DEHKTL	2hmza	24	HHHHHH	HHHHHH	E	4rhv1	217	CCCCEE	CCCCEE	E	100
DMVELQ	1a0i_	191	CHHHHH	CHHHHH	E	1kit_	597	EEEECC	EEEEEC	B	91.7
YDSVID	1ak0_	238	HHHHHH	HHHHHH	E	1ipsa	287	CCCEEC	CCEEEC	E	91.7
VTAMLL	1ble_	77	CEEEEE	EEEEEE	B	1gpmb	240	HHHHHH	HHHHHH	B	91.7
VAAVKA	1broa	149	HHHHHH	HHHHHC	E	1kid_	90	EEEEEC	EEEEEC	B	91.7
LGLVLD	1ceo_	83	CEEEEE	CEEEEE	B	1vnc_	135	HHHHHH	HHHHHC	E	91.7
SKVDDF	1frvb	303	CCCECC	CCCCCC	E	2pgd_	36	HHHHHH	HHHHHH	E	91.7
KHLEAG	1gdlo	107	HHHHCC	HHHHCC	E	2pola	254	EEEEEE	EEEEEC	E	91.7
PAAAAI	1goh_	171	CCEEEE	CCCEEE	B	1tca_	280	HHHHHH	HHHHHH	E	91.7
SAAHAL	1gtra	442	ECCCCE	ECCCCC	E	1qnf_	262	HHHHHH	HHHHHH	E	91.7
LVQFGV	1hava	13	EEEEEE	HEEEEE	B	1ycsb	53	HHHHCC	HHHHCC	E	91.7
GDAIIE	1iso_	407	HHHHHH	HHHHHC	E	1iyv_	66	CCEEEE	CCEEEE	E	91.7
PVIERL	1jer_	75	CEEEEC	CEEEEE	E	1kvu_	42	HHHHHH	HHHHHH	E	91.7
QSFEQV	1maz_	68	HHHHHH	HHHHHH	E	2hft_	110	EEEEEE	EEEEEC	E	91.7
KKGATL	1pkn_	118	CCCCEE	ECCCEE	E	1ycc_	9	HHHHHH	HHHHHH	E	91.7
GSAAVL	1rgs_	184	CCEEEE	CEEEEE	E	1xjo_	101	HHHHHH	HHHHHH	B	91.7
PRQALV	1whtb	134	HHHHHH	HHHHHH	B	3pchm	32	CCCCCE	CCCCCC	E	91.7
IDLLLA	1ako_	228	CEEEEE	EEEEEE	B	2pfkd	272	HHHHHC	HHHHHH	E	83.3
GKLVRD	1asya	364	HHHHHH	HHHHHH	E	1igna	66	CCECEC	CEEEEC	E	83.3
QVKYLG	1ax4a	330	HHHHHH	HHHHHH	B	1mml_	242	CEEECC	CEEEEE	E	83.3
TLQLDV	1bgc_	98	HHHHHH	HHHHHH	E	1fds_	61	EEECCC	EEEEEC	E	83.3
SVVVSG	1bgp_	117	HHHHCC	HHHCCC	E	1tdtc	228	EEEEEE	EEEEEC	B	83.3
LPVIDS	1bib_	83	CCEECC	EEEECC	E	1dkgb	59	HHHHHH	HHHHHH	B	83.3
INLDIP	1cdb_	17	ECCCCC	EEECCC	E	1qapa	11	HHHHHH	HHHHHH	E	83.3
MGGVSE	1chd_	170	CCCCCE	HCCCEE	E	1mhlc	336	HHHHHC	HHHHHC	B	83.3
SGIVSG	1csee	104	HHHHHH	HHHHHH	E	1phc_	384	CCCECE	CCCCCC	B	83.3
LLLAGY	1gotb	283	EEEEEE	EEEEEC	B	2pfkd	274	HHHCCC	HHHHCC	E	83.3
LKLAGR	1itg_	48	HHHHHH	HHHHHC	E	1wba_	103	EEECCC	EEEECC	E	83.3
AELKPL	1mbd_	84	HHHHHH	HHHHHH	E	1vdc_	27	CCCCCE	CCCEEE	E	83.3
VLDAKT	1nox_	2	CCCHHH	CHHHHH	E	2bbkh	284	EEECCC	EEECCC	E	83.3
ELKGTS	1ospo	40	EEEEEE	EEEEEE	E	2abd_	60	CCCCCC	HHCCCC	E	83.3
NLTSVL	1ovab	279	EHHHHH	CHHHHH	E	1ysc_	168	CEEEEE	EEEEEE	E	83.
FFLFDD	1psla	74	HHHHHH	HHHHHH	E	2lgsa	126	EEEECE	EEEEEC	B	83.3
GNVTAE	1qapa	252	CCCCHH	CCCCHH	E	1xsoa	83	EEEEEE	CCEEEE	E	83.3
RPRFER	1rgs_	238	HHHHHH	HHHHHH	E	2kdb_	172	CCCCEE	CCCCCC	E	83.3
KYPVDL	2dri_	260	EEEECC	EEEEEE	E	3pbga	55	HHHHHH	HHHHHH	E	83.3
LTGVKV	1alo_	33	CCCCCC	HCCCEE	B	2dmr_	687	EEEEEE	EEEEEE	B	75
APDEWI	1amp_	87	EEEEEE	CCCEEE	E	1spua	616	CCCCHH	CCCCHH	E	75
LVTEVE	1bip_	97	CCCCCC	HHCCCE	E	1pkn_	176	EEEEEE	EEEEEE	B	75
TGRAAV	1cfe_	63	CHHHHH	CHHHHH	B	1dar_	36	HCCCCE	HCEEEE	B	75
VGAELE	1csee	191	CCCCCC	CCCEEE	E	1esc_	158	HHHHHH	HHHHHH	E	75
RTYKLL	1csn_	52	HHHHHC	HHHHHH	E	1wba_	127	CCEEEE	EEEEEE	B	75
SLKDGV	1irk_	200	HHHHCC	HHHHCC	E	1ytfc	21	EEEEEE	EECCCE	B	75

TABLE V. (Continued)

Sequence	PDB id	Start res[a]	Structure	Prediction	Solv acc[b]	PDB id	Start res[a]	Structure	Prediction	Solv acc[b]	⟨$Q3$⟩ (%)
YQDTAK	1isua	12	CECCCE	HCCCCC	E	1pbn_	10	HHHHHH	HHHHHH	E	75
LIRDHI	1leb_	14	HHHHHH	HHHHHH	E	1pbn_	131	EEEEEE	EEECCC	E	75
AVAKRL	1php_	75	HHHHHH	HHHHHH	E	1vdc_	119	EEECCC	CCCCCC	E	75
VNSLGE	1tlk_	86	EECCEE	ECCCCC	E	2ilk_	86	HHHHHH	HHHHHH	E	75
PELADL	1yaia	32	EEECCC	EEEECC	E	4pgaa	42	CCHHHH	HHHHHH	E	75
KDVEGI	2omf_	281	EEECCC	CCCCCC	E	5csma	183	CCHHHH	CCHHHH	E	75
ELVGPK	1a0b_	16	HHCCHH	HHHCCC	E	2phla	255	EEEEEC	EEEECC	B	66.7
PKLVTE	1agna	328	HHHHHH	HHHHHH	E	1bip_	95	CCCCCC	HHHHCC	E	66.7
IGGAAV	1chma	175	HHHHHH	HHHHHH	B	1uxy_	277	ECCEEE	CCCHHH	B	66.7
LQALDI	1cto_	9	EEECCC	CEEEEE	E	2fsp_	37	HHHHHH	HHHHHH	E	66.7
ATGSDD	1gotb	241	EEEECC	EEECCC	B	4aaha	194	CCCCHH	ECCCCC	E	66.7
LDKYGD	1knya	24	HHHHHH	HHHCCC	E	2dmr_	720	CECCCC	EECCCC	B	66.7
TLVVGG	1liaa	96	HHHHCC	EEEECC	B	1lxta	55	EEEEEE	EEEEEE	B	66.7
APAAAA	1lml_	247	CCCCCC	CCHHHH	E	1tca_	279	HHHHHH	HHHHHH	E	66.7
VIGLLD	1p38_	80	ECCCCE	EEEEEE	B	1tdtc	31	HHHHHH	HHHHHH	B	66.7
IRAALP	1pda_	172	CCEECC	HHHCCC	E	1php_	37	HHHHHH	HHHHHH	E	66.7
GKIEMG	1tum_	38	CCCCCC	CCCCCC	E	2por_	68	EEEEEE	CEECCC	E	66.7
NMLPLL	1a0i_	167	HHHHHH	HHHHHH	B	2mev4	52	CCCCCC	HHHHHC	E	58.3
EKLIEK	1aa3_	14	CCCECC	EEEEEE	E	1xixb	37	HHHHHH	HHHHHH	E	58.3
VGINHG	1ab8a	114	EEEEEE	EEEECC	B	1arv_	98	HHHHHH	HHCCCC	E	58.3
RLKPEI	1avob	14	HHHHHH	HCCHHH	E	1dora	238	CCCCCC	HHCCCE	E	58.3
DVANAV	1aym2	18	CCCCCE	EEEEEE	E	1pkn_	330	HHHHHH	HHHHHH	B	58.3
IATVNE	1cyx_	23	EEEECE	EECCCE	B	1thtb	273	HHHHHH	EEEHHH	E	58.3
MFGYAT	1dar_	578	CCCHHH	HHHHHH	E	1mxa_	116	EEEEEE	EEEECC	B	58.3
TPNILY	1ggga	154	HHHHHH	HHHHHH	B	1quf_	189	CCCCCC	CHHHHH	B	58.3
RHVYGE	1ggta	695	EEEEEE	CEEEEE	B	2pbal	74	HHHHHH	CEECHH	E	58.3
IWNSSV	1kuh_	24	HHHHHC	HHHCCE	E	1vdc_	210	ECCEEE	EEEEEE	E	58.3
PKATSS	1tnra	61	HHHHCC	CCCCCC	E	2bpa2	35	CECCEE	CCCCEE	E	58.3
GHKIKG	1a0b_	57	HHHHHH	CCEEEE	E	1lnh_	1	CCEEEE	CCEEEE	E	50
LQVEIG	1aa0_	11	HHHHHC	EEEEEE	E	1cewi	54	EEEEEE	EEEEEE	B	50
YRALLE	1ad2_	3	CCCCCC	HHHHHH	E	1dhs_	34	HHHHHH	HHHHHH	E	50
EETLVI	1auk_	254	HHEEEE	CCEEEE	E	1awj_	19	CCCCCC	CCEEEE	E	50
VEEVNA	1def_	20	CCCCCC	HHHHHH	E	1gdlo	253	HHHHHH	HHHHHH	E	50
EAGKQA	1dru_	105	HHHHHH	HHHHHH	E	1pbn_	262	CCCCCC	HHHHHH	E	50
PEEVLD	1fds_	195	HHHHHH	CCCCCC	E	3minb	256	CCCCCC	CCCCCC	B	50
YWTYPG	2cba_	188	EEEEEE	EEEECC	B	2mtac	77	CCCCHH	CEECCC	E	50
DLALGK	1ctj_	3	CHHHHH	CCCCCC	E	1smd_	167	EECCCC	CCCCCC	E	41.7
LLPRVA	1efva	100	HHHHHH	CHHHHH	B	1p04a	75	EEEEEE	CCCCCC	E	41.7
SPLAQI	1ak1_	52	HHHHHH	CCCHHH	E	4aaha	27	ECCCCC	CCHHHH	E	33.3
TVGGVT	1ar1a	368	HHHHHH	EEEEEE	B	2wea_	91	EECCEE	EEEEEE	E	33.3
ATVKAK	1prcc	26	HHHHHH	CCCCCC	E	1rgs_	99	CEEEEC	EEEEEE	E	33.3
TLIKDG	1pioa	186	CCHHHH	CEEEEC	B	1spua	25	EEEECC	HHHHCC	B	25
KQIIAN	1ixh_	43	HHHHCC	EEEEEC	E	1shca	127	CEEEEE	HHHHHH	E	8.3
7-mers											
LITTAHA	1cgu_	121	HHHHHHH	HHHHHHH	E	1bgla	833	EEEEEEE	EEEEEEE	B	100
AVLSAIG	1mxa_	90	EEEEEEE	EEEEEEE	E	1tml_	203	HHHHHHC	HHHHHHC	E	100
ASVKQVS	1amp_	63	EEEEEEC	EEEEEEE	E	1gky_	83	HHHHHHH	HHHHHHH	E	92.9
KGLEWVS	1thg_	191	HHHHHHH	HHHHHHH	B	1igmh	43	CCEEEEE	CCCEEEE	E	92.9
GTATHTV	1goh_	577	CEEECCE	EEEEEEE	B	1sly_	500	HHHHHHH	HHHHHHH	E	78.6
EKAYLRT	1pgs_	177	CEEEEEE	HHHEEEE	B	2sblb	699	HHHHHHH	HHHHHHH	B	78.6
RRDALLE	1ayl_	305	CCCCEEE	HHCEEEE	E	1qapa	3	HHHHHHH	CHHHHHH	E	71.4
LRRARAA	1gdlo	194	CCCCCEC	CCCHHHC	B	1pta_	52	HHHHHHH	HHHHHHC	E	71.4
VQNLQVE	1aa0_	8	HHHHHHH	HHEEEEE	E	1ipsa	225	CCCEEEE	CCCEEEE	B	64.3
QEALEIA	1tif_	30	HHHHHHH	HHHHHHH	E	1wtua	66	CCCCCCC	CCEEEEE	E	64.3
VDAELFL	1bcmb	174	EEHHHHH	HHHHHHH	E	1pex_	37	CCCEEEE	CCCCCCH	E	57.1
DLKIQER	1cdb_	99	EEECCCC	HHHHHCC	E	1nre_	32	HHHHHHH	CHHHHHH	E	57.1
ATADFVA	1clc_	264	HHHHHHH	HHHHHHH	B	2mev1	15	CCCCCCC	CEEEEEE	E	57.1
RSSLPGF	1aerb	105	HHHHHHE	CCCCCCC	E	1cto_	88	CCCCCCC	ECCCCCC	E	42.9
LSLAVAG	1bgw_	36	HHHHHHH	CCCEECC	E	1mdah	67	CCEEEEC	CCCEEEE	B	35.7
SNRFYTL	1aym2	51	CCCCEEC	CEEEEEE	E	1pax_	72	HHHHHHH	CCCEEEE	E	21.4
8-mers											
GSLVALGF	1pht_	33	HHHHCCCC	CEEEEEEE	E	1wbc_	72	CCCEEEEE	CCEEEEEEE	B	43.8

[a] Starting position in the protein sequence.
[b] Average solvent accessibility: E, exposed; B, buried.

TABLE VI. Performance of Different Methods on Chameleons

Input	Method[a]		Q3 (%)	
Single sequence	NN	Test 58.9		Train 59.8
	GORIV		55.2	
Multiple sequence (PSI-BLAST)	NN	Test 75.1		Train 76.4
	PHD		73.68	
	PSI-PRED		75.59	
	JPRED		72.94	
	PRED2ARY		73.31	
	DSC		69.89	
	NNSSP		69.80	
	PREDATOR		68.90	

[a] NN are the neural network-based predictors described in this work used in testing and training and available at www.biocomp.unibo.it; GORIV[40]: pbil.ibcp.fr/cgi-bin/npsa_automat.pl; PHD[18]: dodo.cmpc.columbia.edu/predictprotein; PSI-PRED[25]: globin.bio.warwick.ac.uk.psiform.html; JPRED[21]: circinus.ebi.ac.uk:8081/submit.html; PRED2ARY[24]: www.cmpharm.ucsf.edu/~jmc/pred2ary; DSC,[41] NNSSP,[23] and PREDATOR[22] as implemented in JPRED.[21]

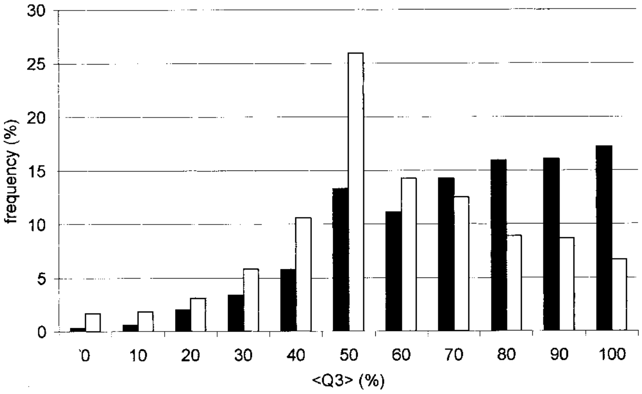

Fig. 2. Bar graph showing the distribution of the ⟨Q3⟩ scores for the 2,576 couples of the database of chameleons predicted using as input single sequence (white bars) and multiple sequence (black bars, obtained with PSI-BLAST).

analyzing the prediction of a small subset of the residues included in the database (about 15%). Indeed, with the same threshold of accuracy only 5% (single sequence) and 9% (multiple sequence) of the proteins of the database are predicted (see Fig. 1).

In Table VII, high-scoring predictions <Q3> ≥ 90%) obtained using multiple sequences are listed by grouping the chameleon couples according to structural types and categories of solvent accessibility. It is evident that the effect of multiple alignment on the prediction is distributed over the most represented structural classes, rather independently of the category of solvent accessibility. Noticeably, the best-predicted sequences are those in the E/H structures, which are not the most abundant structural class in the database of chameleons (see Table III). This is in agreement with the previous finding that multiple sequence comparison techniques are efficient in the alignment of structural regions.[42]

TABLE VII. Analysis of Chameleons Predicted With an Accuracy ≥90%

Segment 1 struct type	Segment 2 struct type	%[a]	% (both exposed)[a]	% (mixed)[a]	% (both buried)[a]
C	E	17.4 (144)	17.6 (85)	13 (46)	30.8 (13)
C	H	37.3 (407)	39.1 (307)	31.4 (86)	35.7 (14)
E	H	65.7 (350)	67.3 (153)	64.2 (148)	65.3 (49)
C	E-H	4.3 (23)	6.3 (16)	0 (7)	— (0)
E	C-H	24.9 (253)	20.1 (134)	29 (100)	36.8 (19)
H	C-E	37 (1011)	35.1 (616)	40.3 (308)	39.1 (87)
C-E	C-H	12.2 (279)	10.8 (186)	13.9 (79)	21.4 (14)
C-E	C-E	4.5 (22)	9.1 (11)	0 (10)	0 (1)
C-E	E-H	10.5 (19)	7.7 (13)	20 (5)	0 (1)
C-H	E-H	0 (6)	0 (4)	0 (2)	— (0)
C-H	C-H	6.7 (30)	7.7 (26)	0 (4)	— (0)
C-E	C-E-H	0 (11)	0 (5)	0 (5)	0 (1)
C-H	C-E-H	11.1 (18)	15.4 (13)	0 (4)	0 (1)
E-H	C-E-H	0 (2)	0 (2)	— (0)	— (0)
C-E-H	C-E-H	0 (1)	0 (1)	— (0)	— (0)

[a] Percentage of chameleon couples predicted with ⟨$Q3$⟩ ≥90% and sorted by structural type and category of solvent accessibility. The corresponding total number of couples in the database is showed within parentheses. The number of couples with ⟨$Q3$⟩ ≥90% is 857, 42 of which are 6- and 7-mers.

CONCLUSIONS

In this study, we show that chameleon sequences can be predicted by methods of secondary structure prediction relying on the information contained in the context of a local window sliding over the protein sequence or including the protein sequence profile computed from the multiple alignment. When single sequence is used, ambiguous mapping during training hampers the discrimination of the different structural types, and the efficiency is lower than that obtained in cross-validation over the whole testing set. Fifteen percent of the couples, however, are quite well discriminated ($<Q3> \geq 90\%$), indicating that in some cases the information included in the 17-residue-long window is sufficient to compute the correct prediction. This suggests that chameleons might be stabilized by the local protein context in a given secondary structure type similarly to other well-predicted segments in the protein sequence.[12] Most importantly, when sequence profiles are used instead of single sequence the inclusion of evolutionary information in the input window is partially sufficient to mitigate the ambiguity associated with the structural classification of chameleon segments. As a matter of fact, the prediction efficiency of chameleons levels that attained on the whole testing set. Our results, in sum, also highlight the generalization capability of the neural network-based predictive methods for those segments, which, in proteins, have the same sequence but different secondary structure.

ACKNOWLEDGMENTS

Financial support to this work was provided by a grant of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) delivered to the project “Structural, Functional and Applicative Prospects of Proteins from Thermophiles.” R.C. was also partially supported by a grant for a target project in Biotechnology from the Italian Centro Nazionale delle Ricerche (CNR).

REFERENCES

1. Socci ND, Onuchic JN, Wolynes PG. Protein folding mechanisms and the multidimensional folding funnel. Proteins 1998;32:136–158.
2. Dill KA. Polymer principles and protein folding. Protein Sci 1999;8:1166–1180.
3. Bellew RM, Sabelko J, Gruebele M. Direct observation of fast protein folding. The initial collapse of apomyoglobin. Proc Natl Acad Sci USA 1996;93:5759–5764.
4. Minor DL, Kim PS. Context dependent secondary structure formation of a designed protein sequence. Nature 1996;380:730–734.
5. Freund SMV, Wong KB, Fersht AR. Initiation sites of protein folding by NMR analysis. Proc Natl Acad Sci USA 1996;93:10600–10603.
6. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–853.
7. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci USA 1996;93:5814–5818.
8. Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc Natl Acad Sci USA 1984;81:1075–1078.
9. Argos P. Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. J Mol Biol 1987;197:331–348.
10. Cohen BI, Presnell SR, Cohen FE. Origins of structural diversity within sequentially identical hexapeptides. Protein Sci 1993;2: 2134–2145.
11. Mezei M. Chameleon sequences in the PDB. Protein Eng 1998;6: 411–414.
12. Sudarsanam S. Structural diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations. Proteins 1998;30:228–231.
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 28;2000:235–242.
14. Compiani M, Fariselli P, Casadio R. Noise and random-like behaviour of perceptrons: theory and applications to protein structure prediction. Phys Rev E 1997;55:7334–7343.
15. Qian N, Sejnowski J. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol. 1988;202:865–884.
16. Holley HL, Karplus M. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci USA 1989;86:152–156.
17. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120.
18. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
19. Rost B, Sander C. 3rd generation prediction of secondary structure. In: Webster DM editor. Predicting protein structure. Totowa, NJ: Humana Press, 2000 (in press).
20. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994; 19, 55–72.
21. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34: 508–519.
22. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure. Proteins 1997;27:329–335.
23. Salomov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest neighbor algorithms and multiple sequence alignments. J Mol Biol 1995;247:11–15.
24. Chandonia JM, Karplus M. New methods for accurate prediction of protein secondary structure. Proteins 1999;35:293–306.
25. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
26. Rost B. Marrying structure and genomics. Structure 1998;6:259–263.
27. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28:45–48.
28. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern of hydrogen—bonded and geometrical features. Biopolymers 1983;22:2577–2637.
29. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229:234–238.
30. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
31. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323: 533–536.
32. Fariselli P, Compiani M, Casadio R. Predicting secondary structures of membrane proteins with neural networks. Eur Biophys J 1993;22: 41–51.
33. Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acid Res 1998;26:313–315.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215: 403–410.
35. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein data base search programs. Nucleic Acids Res 1997;25:3389–3402.
37. Bryant SH; Altschul SF. Statistics of sequence-structure threading. Curr Opin Str Biol 1995;5:236–244.
38. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure of protein secondary structure prediction assessment. Proteins 1999;34:220–223.

39. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delauny tesellation and their application in sequence-structure alignment. Protein Sci1997;6: 1467–1481.
40. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:540–553.
41. King RD, Sternberg MJE. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Prot Sci 1996;5:2298–2310.
42. Barton GJ. Protein sequence alignment techniques. Acta Crystallogr. D Biol Crystallogr 1998;54:1139–1146.

APPENDIX

In this study, the efficiency of the predictors is scored using the statistical indexes defined in the following. The network accuracy is:

$$Q3 = P/N \tag{1A}$$

where P is the total number of correct predictions and N is the total number of possible predictions. The average accuracy (<Q3>) is the Q3 value averaged over the couple. The correlation coefficient C for the structural class s (H, E, and C) is defined as:

$$C(s) = (p(s) * n(s) - u(s) * o(s))/[(p(s) + u(s))(p(s) + o(s)) \times (n(s) + u(s))(n(s) + o(s))]^{1/2} \tag{2A}$$

Where, for each class s, p(s) and n(s) are, respectively, the total number of correct predictions and correctly rejected assignments while u(s) and o(s) are the numbers of under and over predictions. The accuracy for each discriminated structure s is evaluated as:

$$Q(s) = p(s)/[p(s) + u(s)] \tag{3A}$$

Where p(s) and u(s) are the same as in eq. (2A). The probability of correct predictions P(s) is computed as:

$$P(s) = p(s)/[p(s) + o(s)] \tag{4A}$$

Where p(s) and o(s) are the same as in eq. (2A).

The segment-based measure (Sov) of the assessment of protein secondary structure prediction is computed as described in Zemla et al. (1999).[38]