Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 342 (2006) 441–451
BBRC
www.elsevier.com/locate/ybbrc

Amino acid propensities for secondary structures are influenced by the protein structural class

Susan Costantini [a,b,c], Giovanni Colonna [b,c], Angelo M. Facchiano [a,b,*]

[a] *Laboratorio di Bioinformatica e Biologia Computazionale, Istituto di Scienze dell'Alimentazione, CNR, Avellino, Italy*
[b] *Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, Italy*
[c] *Dipartimento di Biochimica e Biofisica, Seconda Università di Napoli, Italy*

Received 19 January 2006
Available online 8 February 2006

Abstract

Amino acid propensities for secondary structures were used since the 1970s, when Chou and Fasman evaluated them within datasets of few tens of proteins and developed a method to predict secondary structure of proteins, still in use despite prediction methods having evolved to very different approaches and higher reliability. Propensity for secondary structures represents an intrinsic property of amino acid, and it is used for generating new algorithms and prediction methods, therefore our work has been aimed to investigate what is the best protein dataset to evaluate the amino acid propensities, either larger but not homogeneous or smaller but homogeneous sets, i.e., all-α, all-β, α–β proteins. As a first analysis, we evaluated amino acid propensities for helix, β-strand, and coil in more than 2000 proteins from the PDBselect dataset. With these propensities, secondary structure predictions performed with a method very similar to that of Chou and Fasman gave us results better than the original one, based on propensities derived from the few tens of X-ray protein structures available in the 1970s. In a refined analysis, we subdivided the PDBselect dataset of proteins in three secondary structural classes, i.e., all-α, all-β, and α–β proteins. For each class, the amino acid propensities for helix, β-strand, and coil have been calculated and used to predict secondary structure elements for proteins belonging to the same class by using resubstitution and jackknife tests. This second round of predictions further improved the results of the first round. Therefore, amino acid propensities for secondary structures became more reliable depending on the degree of homogeneity of the protein dataset used to evaluate them. Indeed, our results indicate also that all algorithms using propensities for secondary structure can be still improved to obtain better predictive results.

Keywords: Amino acid propensities; Structural class of proteins; Protein structure; Secondary structure prediction; Statistical methods

The Anfinsen's experiments in the 1950s suggested that the primary amino acid sequence contains the information that specifies the folded native protein structure [1]. On the basis of this principle, in the 1970s some researchers developed methods to predict the native conformation of proteins from their amino acid sequences, one of the most challenging problems in molecular biology. The Chou and Fasman method [2,3], one of the early prediction methods, was based on a statistical procedure based on assigning conformation potentials, or propensities, to all amino acid residues. Conformation potentials, one for each type of secondary structure, are obtained from statistical analysis of proteins of known secondary structure, as ratio of the fractional occurrence of the residue in secondary structure elements of a given type to the fractional occurrence in all structures. For α-helix and β-strand propensities, each amino acid was classified as former, breaker or indifferent. These properties were used to identify potential α- and β-forming sites, which were then extended along the protein chain as long as the average propensity values calculated over a window of 5 or 6 residues were above a threshold value.

Several other prediction methods were developed over the years, based on different algorithms such as information theory [4], neural networks [5–8], nearest neighbour methods [9], multiple alignments [10–12], combination of

* Corresponding author.
E-mail address: angelo.facchiano@isa.cnr.it (A.M. Facchiano).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.01.159

multiple alignment and neural network [13], and hydrophobicity profiles [14]. These methods have reached relevant improvement in the accuracy of prediction, in comparison to the original Chou and Fasman method. Moreover, the reliability of the Chou and Fasman method has been criticized since the 1980s [15,16]. Nevertheless, many authors still use amino acid propensities or the Chou and Fasman method, for structure predictions [17–24] as well as for evolution studies [25] and in developing or evaluating new prediction methods [26–39]. The large use of this method may be due to its approach, simple but clear when compared to the most recent and accurate, but more sophisticated [40]. In fact, even if the original method has low accuracy in comparison to the most recent approaches, anyway the propensities for the different secondary structures represent intrinsic properties of amino acids and their use in developing new methods became successful [41,42].

The original dataset from which the Chou–Fasman parameters were calculated was quite small: it contained only 15 proteins, consisting of 2473 amino acid residues [2,43]. In 1989, the dataset was extended to include 64 proteins with 11,445 amino acid residues [3]. Later, in the 1998 the size of the dataset was expanded to include 144 proteins in order to analyse the reliability of the Chou–Fasman parameters [16].

The concept of protein structural classes was introduced by Levitt and Chothia [44] based on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins. According to this concept, protein folds can be classified into one of four classes: all-α, all-β, α/β, and α + β. Since then, various quantitative classification rules have been proposed based on the percentages of α-helices and β-sheets in a protein [6,45–47].

In our work, we examined the problem to verify how the protein dataset used to compute amino acid propensities can affect the results, in particular, by using either large but not homogeneous datasets or smaller but homogeneous datasets consisting of only all-α, only all-β, or only α–β proteins. We have calculated the amino acid propensities for three types of secondary structures for 2168 proteins, i.e., the whole PDBselect dataset. Then, we subdivided these proteins in three secondary structural classes and calculated the amino acid propensities for each class. The prediction of secondary structure has been made using the different amino acid propensities calculated. Results are compared and discussed to evaluate the better criteria to choose protein dataset for computing amino acid propensities.

Methods

Database and definition of protein secondary structure. All analyses were performed using PDBselect [48] as a set of experimentally determined, non-redundant protein structures in the Protein Data Bank (see http://homepages.fh-giessen.de/~hg12640/pdbselect). We used the PDBselect list with <25% sequence homology, released in December 2003, which contained 2216 protein chains.

The secondary structure for every PDB entry was assigned by the DSSP algorithm [49] based on the analysis of backbone dihedral angles and hydrogen bonds. DSSP assigns seven different secondary structures, i.e., H: α-helix, G: 3_{10} helix, I: π-helix, E: extended strand, B: residue in isolated β-bridge, S: bend, and T: H-bonded turn. In addition, a "coil" state is assigned when no secondary structure is recognized. We applied the convention to define H, G, I as helix, E and B as strand, and others as coil [50,51]. 2168 of 2216 PDBselect proteins were accepted by DSSP for the analysis and constituted the total PDB subset for our work.

Assignment of structural class. The secondary structural content has been used to assign the protein secondary structural class, according to two different definitions of structural classifications. Nakashima et al. [45] consider proteins with >15% α-helical content and <10% β-strand content as all-α proteins, with <15% α-content and >10% β-content as all-β proteins, with >15% α-content and >10% β-content as mixed proteins, and the remaining as irregular.

According to the criterion of Chou [46], all-α proteins have at least 40% α-helical content and <5% β-strand content; all-β proteins have at least 40% β-strand content and <5% α-helical content; mixed proteins (considering the combination of α + β and α/β classes) contain more than 15% α-helical and 15% β-strand contents; irregular proteins have <10% α-helical and β-strand contents.

Propensity of amino acids in different secondary structural types. The residue propensity values in different secondary structural types (P_{ij}) were determined from the ratio of the residue's frequency of occurrence in helices, β-strand, and coil versus its frequency of occurrence in the protein subset:

$$P_{ij} = \frac{(n_{ij}/n_i)}{(N_j/N_T)},$$

where n_{ij} is the number of residues of type i in structure of type j, n_i is the total number of residues of type i, N_j is the total number of residues in structure of type j, and N_T is the total number of residues in the subset of PDB used in this analysis. Then, these values of propensities have been normalized as follows:

$$P_{ik}^{norm} = \frac{(P_{ik} - P_k^{min})}{(P_k^{max} - P_k^{min})},$$

where P_{ik} is the propensity of each amino acid in secondary structure element of type k (α, β or coil), P_k^{min} and P_k^{max} are the minimum and maximum values between the propensities P_{ik}.

Prediction of secondary structure. Starting from the N-terminal of each protein sequence, a running window of n amino acids is taken. The average value of α-helical propensities $\langle P_\alpha \rangle$, β-strand propensities $\langle P_\beta \rangle$, and coil propensities $\langle P_c \rangle$ has been determined for the n amino acids of each segment. These propensities have been determined by using windows of different lengths for three secondary structure elements (w_H, w_E, and w_c) and multiplied by different coefficients ($coeff_H$, $coeff_E$, and $coeff_c$). An exaustive scan for different windows and coefficients was made in order to find the values giving the better results.

The predicted secondary structure for the middle amino acid in the examined segment was assigned by choosing the higher value between the three average propensities of the segment. In this manner, if $\langle P_\alpha \rangle$ for one segment of length 7 is higher than $\langle P_\beta \rangle$ and $\langle P_c \rangle$, it has been assigned the secondary structure of type "H" to the 4th amino acid of that sequence. This procedure has been repeated for all proteins collected in the PDBselect database.

The prediction quality was examined by both resubstitution and jackknife tests.

Resubstitution test. The so-called resubstitution test is an examination for the self-consistency of a prediction algorithm. When the resubstitution test is performed for the current study, the secondary structure elements of each protein in a given dataset are predicted by the propensities derived from the same dataset, the so-called training dataset.

As a consequence, the propensities derived from the training dataset include the information of the protein used in the test. This will certainly give a somewhat optimistic error estimate because the same proteins are

used to derive the propensity values necessary for the predictions and to test themselves.

Jackknife test. This analysis is also called leave-one-out test [47] in which each protein in the PDBselect dataset is singled out in turn as a test sample and the propensity values are calculated from training all the remaining proteins (without using this protein). In other words, the secondary structure elements of each protein are predicted by the propensities derived using all other proteins except the one that is being predicted. During the process of jackknife analysis, each protein has one chance to be the test sample, and for other tests this protein will be included in the training dataset [52,53]. Indeed, for this test, the same windows and coefficients optimized in the resubstitution test have been used.

Evaluation of predictive accuracy. To evaluate the success of any prediction, it was necessary to compare the predicted secondary structure elements for each residue of a protein with the assignments made by the DSSP algorithm [49]. The percentage of residues predicted correctly in the conformational state k is given by:

$$Q_k = \%_k = 100 \cdot \frac{n_k}{N_k},$$

where k represents the α-, β- and coil regions in the native protein structure as determined experimentally, n_k is the number of correctly predicted residues in the state k, and N_k is the total number of residues in the conformational state k in the protein.

The percentage of total residues in the protein identified correctly in helices, β-strand and coil is

$$Q_3 = \%_N = 100 \cdot \frac{(N_T - N_x)}{N_T},$$

where N_T is the total number of residues in the protein and N_x is the total number of incorrectly predicted residues in the protein.

Comparison to Chou–Fasman predictions. Chou–Fasman predictions have been made by using the PeptideStructure program [54]. PeptideStructure uses the original method of Chou and Fasman [55,56] to predict helices, sheets, and turns. It resolves overlapping regions of α-helices and β-sheets with the overall probability procedure introduced by Nishikawa [57]. In the PeptideStructure program, the Chou–Fasman rules are slightly modified as follows. For predicting helix the conditions, that in a segment of four amino acids $\langle P_\alpha \rangle$ must be greater than 1.0 and greater than $\langle P_\beta \rangle$, are not used; but the condition of a minimum length of five residues is conserved in order to predict sheet. PeptideStructure was written by Dr. B.J. Foertsch and G. Herrmann, and was modified for compatibility with Version 5 of the Wisconsin Package by John Devereux. The program has been described and used by many authors [58–61].

Results

Analysis of PDBselect as a unique set

The PDBselect release of December 2003 included 2216 structures having homology percentage <25%. We assigned the secondary structure for 2168 proteins by using the DSSP program (the others report only α carbons and DSSP did not assign the secondary structure). We simplified the 8-state secondary structure as a three-state secondary structure, considering H, G, and I as helix, B and E as β structure, and the others as coil (see Methods for details about the 8 states).

We calculated the composition in amino acids for the whole dataset. Results are shown in Table 1. Then we calculated the amino acid propensities for all proteins in three types of secondary structures (helix, β-strand, and coil). Results are also shown in Table 1. We observe that Ala, Glu, Lys, Leu, Met, Gln, and Arg have P_α higher than the other propensities; Cys, Phe, Ile, Thr, Val, Trp, and Tyr have P_β higher than others and Gly, Asn, Asp, Pro, His, and Ser have P_c higher than others.

Table 1
Frequency of amino acids and their propensities for the three secondary structure states in the PDBselect dataset

Amino acid	Frequency	P_α	P_β	P_c
A—Ala	7.73	1.39	0.75	0.8
C—Cys	1.84	0.74	1.31	1.05
D—Asp	5.82	0.89	0.55	1.33
E—Glu	6.61	1.35	0.72	0.86
F—Phe	4.05	1.01	1.43	0.76
G—Gly	7.11	0.47	0.65	1.62
H—His	2.35	0.92	0.99	1.07
I—Ile	5.66	1.04	1.71	0.59
K—Lys	6.27	1.11	0.83	1
L—Leu	8.83	1.32	1.1	0.68
M—Met	2.08	1.21	0.99	0.83
N—Asn	4.5	0.77	0.62	1.39
P—Pro	4.52	0.5	0.44	1.72
Q—Gln	3.94	1.29	0.76	0.89
R—Arg	5.03	1.17	0.91	0.91
S—Ser	6.13	0.82	0.85	1.24
T—Thr	5.53	0.76	1.23	1.07
V—Val	6.91	0.89	1.86	0.64
W—Trp	1.51	1.06	1.3	0.79
Y—Tyr	3.54	0.95	1.5	0.78

Evaluation of propensities reliability by secondary structure prediction

On the basis of these propensity values we predicted the secondary structure elements for all proteins (as described in the Methods) and compared the results to predictions obtained by using the Chou and Fasman method. All the predictions made were evaluated by comparison to the secondary structure assigned with the DSSP program. Results are shown in Table 4A, while all windows and coefficient used are reported in the Supplementary Table. In the resubstitution test, the percentage of total correct predictions is of 56.7%. In the jackknife test, the percentages of correct predictions, compared with those by resubstitution test, are decreased by 0.2% for β-strand ($Q_\beta = 52.5\%$) and increased by 0.1% for helices ($Q_\alpha = 60.1\%$).

For a comparison, we predicted the secondary structure of the proteins in the dataset with the method of Chou and Fasman. This method predicted correctly 51.9% of secondary structure elements (see Table 4A).

Analysis of PDBselect subsets: (i) structural classification according to Nakashima et al. [45]

The 2168 proteins from the PDBselect dataset were subdivided in three classes, i.e., all-α, all-β, and α–β, according to two definitions of structural classification (see Methods for details). On the basis of Nakashima's criterion [45], 2091 proteins were classified as follows: 627 as all-α, 552 as all-β, and 912 as α–β. The remaining proteins from the

PDBselect dataset did not satisfy the requirements for classification in any of these three classes.

The amino acid compositions of each class, as well as for the whole 2091 proteins set, are shown in Table 2. Comparing the occurrence of each amino acid in all three classes, it can be observed that Ala, Leu, Glu, Lys, Gln, and Arg are more present in all-α proteins. Cys, Gly, Asn, Pro, Ser, Thr, Val, Trp, and Tyr are more present in all-β proteins. In all-α class, the most frequent residues are Ala, Glu, Lys, and Leu. In all-β class, the most frequent residues are Gly, Leu, Ser, Thr, and Val. In α–β class, the most frequent residues are Ala, Glu, Gly, Leu, and Val. The residue composition in the whole 2091 proteins set is similar to that in α–β class.

The residue propensities in three types of secondary structure for the proteins in the three classes have been calculated as described in the Methods and are also reported in Table 2.

The propensities of the 2091 proteins dataset are very similar to the propensities evaluated with the PDBselect dataset (reported Table 1), as expectable being the latter composed by the same proteins with the addition of other 77 proteins. Very similar are also the results obtained for the whole dataset and the α–β proteins dataset. On the contrary, differences exist by comparing the whole dataset with the all-α protein dataset, as well as the whole with the all-β proteins dataset and the all-α with the all-β dataset. A more detailed description of these differences is reported in Discussion.

Table 2
Analysis of amino acid properties in protein datasets classified according to Nakashima [45]

(A)—Frequency of amino acids

Amino acid	All-α	All-β	α–β	All
A—Ala	8.81	6.2	8.00	7.74
C—Cys	1.75	2.75	1.4	1.78
D—Asp	5.62	5.79	5.9	5.8
E—Glu	7.26	5.69	6.76	6.6
F—Phe	3.91	4.07	4.11	4.05
G—Gly	5.76	8.06	7.2	7.07
H—His	2.45	2.15	2.41	2.35
I—Ile	5.48	5.26	5.92	5.66
K—Lys	6.82	6.1	6.12	6.25
L—Leu	10.44	7.11	8.97	8.84
M—Met	2.46	1.68	2.12	2.09
N—Asn	4.28	5.14	4.33	4.5
P—Pro	3.9	5.19	4.46	4.5
Q—Gln	4.74	3.75	3.73	3.95
R—Arg	5.42	4.41	5.12	5.01
S—Ser	5.77	7.12	5.85	6.11
T—Thr	4.87	6.59	5.35	5.52
V—Val	5.69	7.39	7.23	6.92
W—Trp	1.44	1.77	1.41	1.5
Y—Tyr	3.15	3.77	3.6	3.54

(B)—Amino acid propensities for secondary structures

Amino acid	All-α			All-β			α–β			All		
	P_α	P_β	P_c	P_α	P_β	P_c	P_α	P_β	P_c	P_α	P_β	P_c
A—Ala	1.18	0.59	0.71	1.53	0.91	0.98	1.39	0.75	0.78	1.39	0.75	0.8
C—Cys	0.9	0.95	1.17	0.81	1.12	0.93	0.79	1.36	0.99	0.75	1.34	1.03
D—Asp	0.85	0.65	1.28	1.23	0.55	1.34	0.9	0.54	1.35	0.89	0.55	1.34
E—Glu	1.17	0.5	0.75	1.66	0.86	1.01	1.37	0.7	0.84	1.35	0.72	0.85
F—Phe	1.06	1.44	0.87	0.89	1.42	0.67	0.99	1.42	0.77	1	1.43	0.76
G—Gly	0.57	0.68	1.77	0.49	0.55	1.46	0.47	0.67	1.66	0.48	0.66	1.63
H—His	0.92	1.13	1.13	0.99	0.98	1.02	0.89	1.02	1.09	0.92	0.98	1.07
I—Ile	1.14	1.91	0.68	0.81	1.56	0.56	0.98	1.81	0.57	1.04	1.7	0.59
K—Lys	1.02	0.86	0.98	1.13	0.94	1.02	1.14	0.78	1	1.11	0.83	1
L—Leu	1.18	0.97	0.69	1.28	1.29	0.71	1.28	1.12	0.68	1.32	1.09	0.68
M—Met	1.08	1.11	0.85	1.25	1.1	0.87	1.17	1.05	0.82	1.21	0.99	0.83
N—Asn	0.79	0.74	1.38	0.87	0.61	1.35	0.79	0.59	1.41	0.77	0.62	1.4
P—Pro	0.5	0.95	1.86	0.64	0.35	1.6	0.53	0.45	1.73	0.5	0.44	1.73
Q—Gln	1.13	0.52	0.81	1.48	0.91	0.99	1.29	0.72	0.89	1.28	0.76	0.89
R—Arg	1.09	0.84	0.86	1.16	1.05	0.93	1.17	0.87	0.92	1.17	0.9	0.9
S—Ser	0.83	0.75	1.31	0.93	0.83	1.16	0.85	0.81	1.24	0.82	0.85	1.24
T—Thr	0.87	1.51	1.18	0.62	1.11	0.98	0.79	1.2	1.08	0.76	1.23	1.08
V—Val	1.1	2.1	0.75	0.7	1.59	0.56	0.85	1.9	0.64	0.88	1.85	0.63
W—Trp	1.1	1.15	0.82	1.09	1.34	0.7	1.1	1.17	0.82	1.06	1.3	0.79
Y—Tyr	1.06	1.66	0.84	0.87	1.47	0.63	0.93	1.42	0.83	0.95	1.5	0.78

Evaluation of propensities reliability by secondary structure prediction

The propensities calculated for each class of proteins have been used to predict secondary structure elements for proteins belonging to the same class. By using the resubstitution test, the percentages of correct predictions (shown in Table 4B) for all-α, all-β, and α–β classes were 62.2%, 57.1%, and 57.0%, respectively. Considering the propensities calculated for all 2091 proteins classified according to Nakashima [45], 56.3% of secondary structure elements have been predicted correctly. Windows and coefficients applied (see Methods) are reported in the Supplementary Table.

The jackknife test has also been used in order to evaluate the prediction quality in each class. For example, for the proteins of all-α class, the secondary structure elements of each protein belonging to this class are predicted by the propensities derived using all other proteins, classified as all-α, except the one that is being predicted. The results of jackknife test obtained for each class are given in Table 4B. The rates of correct prediction by jackknife test are decreased, compared with those by resubstitution test, of 0.1% for Q_3 ($Q_3 = 62.0\%$) in all-α and all-β classes, with differences of 0–0.3% at level of Q_α, Q_β, and Q_{coil}. The percentages of correct predictions for the proteins in α–β class and all 2091 proteins were unchanged with differences of

Table 3
Analysis of amino acid properties in protein datasets classified according to Chou [46]

(A)—Frequency of amino acids

Amino acid	All-α	All-β	α–β	All
A—Ala	8.84	6.09	7.75	7.8
C—Cys	1.5	2.02	1.37	1.47
D—Asp	5.54	5.58	5.91	5.77
E—Glu	7.47	5.57	6.8	6.8
F—Phe	3.94	4.07	4.11	4.06
G—Gly	5.3	7.89	7.2	6.8
H—His	2.42	2.17	2.4	2.37
I—Ile	5.68	5.45	6.03	5.86
K—Lys	6.85	5.86	6.2	6.31
L—Leu	10.89	7.06	8.92	9.16
M—Met	2.53	1.59	2.15	2.17
N—Asn	4.25	5.37	4.31	4.41
P—Pro	3.49	4.84	4.4	4.22
Q—Gln	4.97	3.76	3.75	4.04
R—Arg	5.44	4.16	5.05	5.03
S—Ser	5.77	7.67	5.9	6.06
T—Thr	4.78	7.14	5.41	5.44
V—Val	5.71	7.69	7.36	6.97
W—Trp	1.52	2.04	1.39	1.49
Y—Tyr	3.11	4	3.59	3.51

(B)—Amino acid propensities for secondary structures

Amino acid	All-α			All-β			α–β			All		
	P_α	P_β	P_c	P_α	P_β	P_c	P_α	P_β	P_c	P_α	P_β	P_c
A—Ala	1.15	0.43	0.69	1.59	0.99	0.97	1.41	0.77	0.8	1.35	0.76	0.79
C—Cys	0.98	1.22	1.04	0.47	1.2	0.81	0.76	1.39	0.96	0.83	1.38	0.95
D—Asp	0.84	0.76	1.35	1.71	0.58	1.44	0.91	0.54	1.36	0.88	0.56	1.37
E—Glu	1.13	0.27	0.74	1.66	0.87	1.09	1.38	0.72	0.85	1.32	0.71	0.85
F—Phe	1.07	1.67	0.83	0.68	1.36	0.6	0.99	1.41	0.75	1.01	1.41	0.75
G—Gly	0.62	0.66	1.83	0.52	0.62	1.49	0.47	0.66	1.66	0.49	0.7	1.69
H—His	0.92	1.14	1.17	1.47	0.98	0.99	0.89	1.02	1.09	0.92	0.99	1.09
I—Ile	1.15	1.84	0.66	0.72	1.44	0.51	0.96	1.76	0.56	1.03	1.68	0.57
K—Lys	1.01	0.96	0.98	1.29	0.96	1.02	1.15	0.8	1.01	1.11	0.81	1
L—Leu	1.14	0.91	0.69	1.08	1.29	0.65	1.29	1.11	0.69	1.29	1.05	0.68
M—Met	1.04	1.66	0.89	1.27	1.04	0.93	1.19	1.02	0.83	1.18	0.96	0.84
N—Asn	0.79	0.86	1.46	1.17	0.62	1.43	0.8	0.59	1.43	0.78	0.63	1.44
P—Pro	0.53	1	2.01	0.94	0.34	1.79	0.52	0.45	1.75	0.5	0.45	1.82
Q—Gln	1.11	0.59	0.78	1.24	0.97	1.02	1.33	0.75	0.88	1.29	0.75	0.86
R—Arg	1.07	0.74	0.86	0.92	1.09	0.9	1.17	0.9	0.92	1.16	0.9	0.9
S—Ser	0.84	0.86	1.35	1.04	0.85	1.18	0.84	0.8	1.26	0.81	0.86	1.27
T—Thr	0.89	1.53	1.21	0.58	1.08	0.95	0.79	1.22	1.04	0.78	1.27	1.06
V—Val	1.11	1.95	0.73	0.46	1.46	0.5	0.83	1.84	0.61	0.88	1.85	0.62
W—Trp	1.08	1.21	0.83	0.68	1.25	0.73	1.08	1.2	0.81	1.04	1.27	0.8
Y—Tyr	1.07	1.64	0.83	0.8	1.4	0.54	0.91	1.41	0.81	0.94	1.48	0.78

0.1% for Q_α in α–β class and for Q_α and Q_β in all 2091 proteins.

For a comparison, we predicted the secondary structure of the proteins in the datasets with the method of Chou and Fasman (see Table 4B). The results indicate that 52.1, 51.1, and 52.2% of secondary structure elements have been predicted correctly for proteins of all-α, all-β, and α–β classes, respectively. The total percentage of correct predictions is of 51.9% for all proteins classified according to Nakashima. All these values are lower than that obtained by our method, with the exception of the Q_α in the all-β class.

Analysis of PDBselect subsets: (ii) structural classification according to Chou

The 2168 proteins from the PDBselect dataset were subdivided in three classes, i.e., all-α, all-β, and α–β, on the basis of the criterion of Chou [46], which recognized 1333 proteins as follows: 470 proteins as all-α, 167 proteins as all-β, and 696 as α–β. The remaining proteins from the PDBselect dataset did not satisfy the requirements for classification in any of these three classes.

The amino acid composition in the different class datasets and the residue propensities for the three secondary structure types for the three classes were evaluated as for the Nakashima datasets and are shown in Table 3. Despite little differences in the values, both amino acid composition and propensities for secondary structures were very similar to those of proteins classified according to Nakashima (see Table 2).

Evaluation of propensities reliability by secondary structure prediction

By using the resubstitution test, the propensities calculated for proteins in the datasets derived by the Chou classification have been used to predict secondary structure elements for proteins belonging to the same class. Results are shown in Table 4C. Windows and coefficients applied (see Methods) are reported in the Supplementary Table. For proteins of all-α, all-β, and α–β classes, the percentages of correct predictions were 62.2, 61.3, and 57.2, respectively. Considering the propensities calculated for the whole set of 1333 proteins, classified according to Chou, 57.4% of secondary structure elements have been predicted correctly.

The results of jackknife test obtained for each class are also given in Table 4C. The correct predictions by jackknife test show subtle decreases, compared with those by resubstitution test, but more evident decrease concerns β-strand prediction in all-α class (3.1%) and α helix prediction in all-β class (2.6%). Considering all 1333 proteins the percentages of correct prediction are decreased of 0.7 for Q_α, 1% for Q_β, and 0.2% for Q_3, and increased of 1.2% for Q_{coil}.

For a comparison, the Chou and Fasman method has been used to predict the secondary structures for proteins of the datasets (see Table 4C). The percentages of secondary structure elements predicted correctly are 52.1, 50.9, and 52.3% for proteins of all-α, all-β, and α–β class, respectively, and 52.1% for all 1333 proteins. These values are lower than those obtained by our method, with few exceptions in the Q values of the specific secondary structures (see Discussion).

Table 4
Accuracy of secondary structure prediction in all proteins of PDBselect dataset (A), in protein classes classified according to Nakashima (B), in proteins, classified according to Chou [46] (C), by using the resubstitution test, the jackknife test, and the Chou and Fasman method

(A)—All proteins of PDB select (2168 proteins) Method	Q_α	Q_β	Q_{coil}	Q_3
(This work) Resubstitution test	60.0	52.7	55.9	56.7
(This work) Jackknife test	60.1	52.5	55.9	56.7
Chou and Fasman	55.3	48.2	50.9	51.9
(B)—Proteins classified according to Nakashima Classes	Q_α	Q_β	Q_{coil}	Q_3
All-α				
Resubstitution test	69.7	47.3	49.2	62.2
Jackknife test	69.5	47.0	49.2	62.0
Chou and Fasman	54.7	45.3	47.7	52.1
All-β				
Resubstitution test	44.2	58.0	58.6	57.1
Jackknife test	43.9	58.0	58.6	57.0
Chou and Fasman	47.2	46.8	55.3	51.1
α–β				
Resubstitution test	60.2	55.1	55.2	57.0
Jackknife test	60.1	55.1	55.2	57.0
Chou and Fasman	56.5	49.4	49.8	52.2
2091 proteins				
Resubstitution test	59.7	55.7	53.8	56.3
Jackknife test	59.6	55.6	53.8	56.3
Chou and Fasman	51.9	55.4	48.2	50.9
(C)—Proteins classified according to Chou Classes	Q_α	Q_β	Q_{coil}	Q_3
All-α				
Resubstitution test	70.5	47.0	43.1	62.2
Jackknife test	70.5	43.9	42.7	62.1
Chou and Fasman	54.6	41.8	46.3	52.1
All-β				
Resubstitution test	35.9	66.3	57.8	61.3
Jackknife test	33.3	66.0	57.4	61.0
Chou and Fasman	40.2	46.2	57.3	50.9
α–β				
Resubstitution test	59.2	55.3	56.5	57.2
Jackknife test	59.1	55.3	56.5	57.1
Chou and Fasman	57.0	49.2	50.1	52.3
1333 proteins				
Resubstitution test	60.8	54.4	55.5	57.4
Jackknife test	60.1	53.4	56.7	57.2
Chou and Fasman	55.8	48.4	50.4	52.1

Q_α, Q_β, and Q_{coil} are the percentages of residues predicted correctly in helices, β-strand, and coil, respectively; Q_3 is the percentage of total residues in proteins predicted correctly.

Discussion

We calculated the amino acid propensities in helix, β-strand, and coil for all proteins in the PDBselect dataset

and evaluated their reliability by using them to predict the secondary structure of proteins. The quality of these predictions was examined by resubstitution and jackknife tests. Results obtained with the two tests are in general very similar (differences of 0.1–0.2%), and in particular when the number of proteins in the dataset was higher. This may reflect the fact that in the jackknife test the protein under prediction is excluded from the set when the propensities are evaluated, but this may affect the results in a sensible manner only when the number of proteins in the set is low. The percentage of correct predictions improved of 4.8% the value obtained with the Chou and Fasman method (see Table 4A). This better result may be a consequence of the larger dataset of proteins we used to calculate the amino acid propensities in the three different types of secondary structure, in comparison to the Chou and Fasman work. Better results were already obtained when, in the past years, larger sets of proteins than the original Chou and Fasman article were used [3,16].

Then, we subdivided these proteins in three secondary structural classes according to two definitions [45,46]. For each class, the residue propensities to the three secondary structure types have been calculated and used to predict secondary structure for proteins of that class. This second round of predictions improves the results obtained in the first round, when we used the propensities evaluated on the whole PDBselect dataset. The percentage of correct predictions were similar with both resubstitution and jackknife tests, with differences of 0.1–0.6%. In general, similar results are obtained with the two classifications,

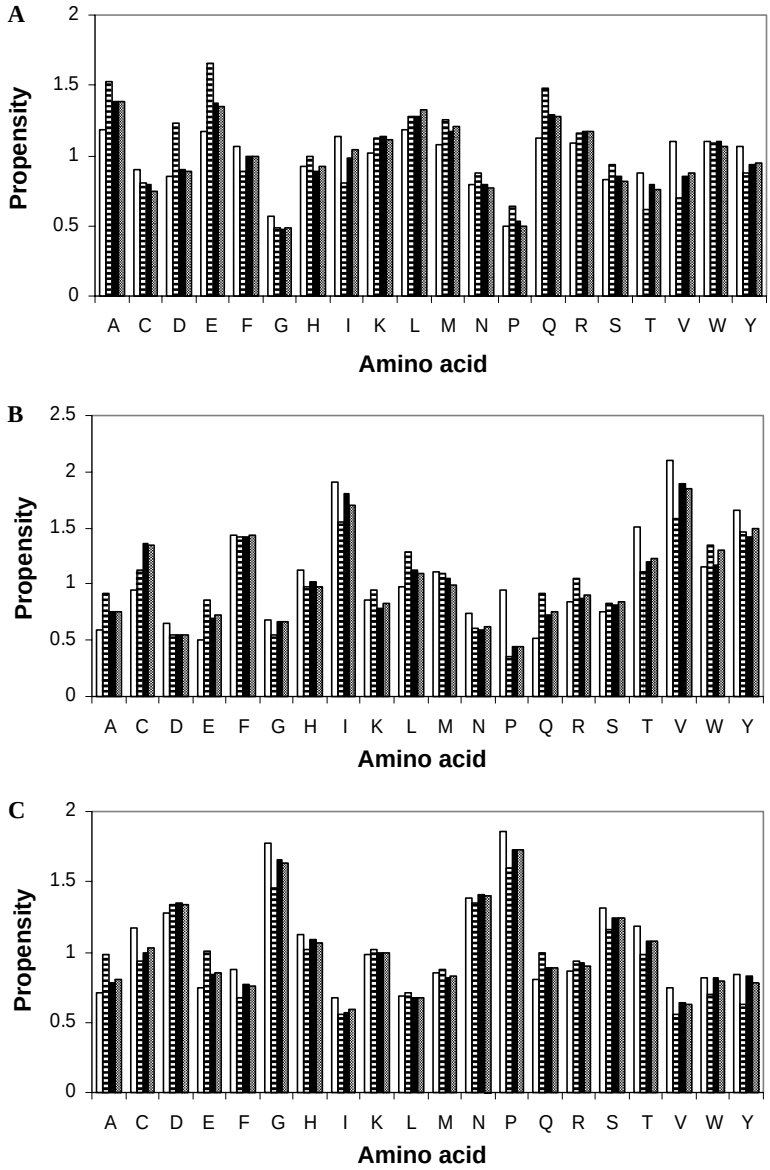

Fig. 1. Bar graphs of the amino acid propensities in proteins classified according to Nakashima [45]. (A) α-helix propensities; (B) β-structure propensities; (C) coil propensities. White bars refer to all-α proteins, horizontal line bars refer to all-β proteins, black bars refer to α–β proteins, and dot bars refer to the whole protein set.

but in some cases the predictions are better for data sets created according to the Chou classification, which assign lower number of proteins to each class (see Table 4B and C). This may appear in contrast with the finding that propensities evaluated for larger datasets give better predictions. Therefore, a possible explanation for the improvement of the results can be related to the more selective rules of Chou's classification, which create more homogeneous datasets.

The prediction accuracies of different methods have been recently compared [62] and better results are obtained by the new methods developed over the years. For any method, the predictions became more accurate for proteins containing α helices but not β structure, probably because these methods are influenced by the effect of neighbouring residues. Our work confirms this finding. Thus, all methods were more reliable in predicting structures having a predominant role of short and medium range interactions among amino acids, which are observed for α helices, while long range interactions influence β structures. The development of prediction techniques specific for each structural class has been considered as a possible way to improve the accuracy of predictions. This might be merged with methods proposed to predict the structural class of a protein by its amino acid sequence [63,64].

It is interesting to note that an improvement of the Chou and Fasman results is obtained with amino acid propensities evaluated with the largest dataset, but also a further improvement is obtained with the smaller but homogeneous datasets after the subdivision according to the

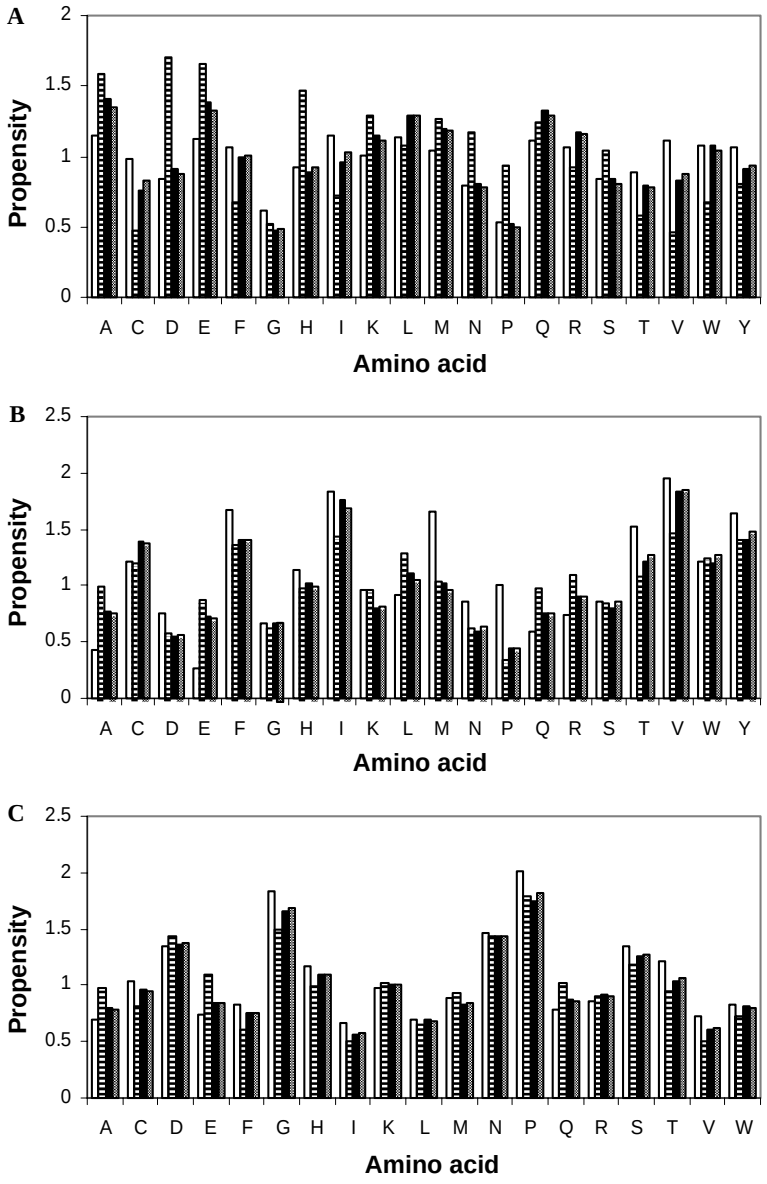

Fig. 2. Bar graphs of the amino acid propensities in proteins classified according to Chou [46]. (A) α-helix propensities; (B) β-structure propensities; (C) coil propensities. White bars refer to all-α proteins, horizontal line bars refer to all-β proteins, black bars refer to α–β proteins, and dotted bars refer to the whole protein set.

structural classification. This means that, although the predictions are improved by increasing the number of proteins in the dataset used to evaluate amino acid propensities, better results can be obtained by using a dataset structurally homogeneous with protein under prediction. Therefore, if the structural class for a given protein can be assigned, as an example by predictive methods, homology evaluations, or experimental results as circular dichroism, it is possible to apply the secondary structure amino acid propensities evaluated for the right class, having better results than other statistical methods based on propensities calculated on mixed sets of proteins.

The analysis of the most populated but non-homogeneous datasets gave us very similar values of amino acid composition and propensities, as appeared by comparing the PDBselect dataset (2168 proteins), the whole sets of proteins classified by Nakashima (2091 proteins) and Chou (1333 proteins), and the α–β class according to both Nakashima and Chou classifications. On the contrary, the analysis of the smaller but homogeneous datasets gave more different values, with propensities in all-α and all-β classes differing by the values obtained in the other sets. Propensities in α–β class appear as mean properties between that in α and β classes, as a consequence of the presence of both secondary structures to a comparable extent. This indicates that the intrinsic propensity of amino acids for secondary structure is influenced by the context of the sequence and structural organization. This aspect could suggest that propensity for secondary structure may not be considered a really intrinsic property of each amino acid, but it must be viewed as influenced by the contest. In any case, the knowledge of the different propensities in different contexts may be useful for the application of these properties in the most suitable manner.

In Fig. 1, we report as bar graphs the comparison of amino acid propensities for each secondary structure type in the three protein structural classes classified according to Nakashima. The most evident differences for the α-helix propensity (Fig. 1A) concern the higher values in all-α class than in all-β class for cysteine (C), phenylalanine (F), isoleucine (I), arginine (R), threonine (T), valine (V), tryptophane (W), and tyrosine (Y). On the contrary, the α-helix propensity is higher in all-β class for alanine (A), aspartate (D), glutamate (E), histidine (H), lysine (K), asparagine (N), proline (P), and serine (S). The same differences appear more evident in the case of data sets created according to the Chou classification (see Fig. 2). Similarly, the comparison of β strand and coil propensities evidences differences in the all-α and all-β classes.

In conclusion, the number of proteins used for the evaluation of amino acid propensities for secondary structure is relevant to improve the reliability of the evaluation, but homogeneous sets of proteins, although smaller, can give even more better results. As a consequence, secondary structure predictions based on the amino acid propensities, introduced in the 1970s, can be improved by increasing the number of proteins used to compute the propensities, but also by using a smaller set which well represents the same structural class of the protein under prediction. Although other predictive approaches exist and give results better than the statistical methods, our results indicate that improvements of statistical methods are still possible. How to improve the composition of the set, in terms of homogeneity degree with the protein under prediction and right number of protein elements, will be the object of our further studies.

Acknowledgments

This work was partially supported by MIUR-FIRB project (Grant RBNE0157EH_003) and by Rete di Spettrometria di Massa (contract FERS n. 94.05.09.103, ARINCO N. 94.IT.16.028). Ph.D. fellowship of Dr. Susan Costantini is supported by E.U.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.01.159.

References

[1] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–230.
[2] P.Y. Chou, G.D. Fasman, Prediction of protein conformation, Biochemistry 13 (1974) 222–245.
[3] P.Y. Chou, Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York: Fasman GD, 1989, p. 549.
[4] J. Garnier, D.J. Osguthorpe, B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, J. Mol. Biol. 120 (1978) 97–120.
[5] N. Qian, T.J. Sejnowski, Predicting the secondary structure of globular proteins using neural network models, J. Mol. Biol. 202 (1988) 865–884.
[6] D.G. Kneller, F.E. Cohen, R. Langridge, Improvements in protein secondary structure prediction by enhanced neural networks, J. Mol. Biol. 214 (1990) 171–182.
[7] B. Rost, C. Sander, R. Schneider, PHD-an automatic mail server for protein secondary structure prediction, Comput. Appl. Biosci. 10 (1994) 53–60.
[8] J.M. Chandonia, M. Karplus, The importance of larger data sets for protein secondary structure prediction with neural networks, Protein Sci. 5 (1996) 768–774.
[9] T.M. Yi, E.S. Lander, Protein structure prediction using nearest-neighbour methods, J. Mol. Biol. 232 (1993) 1117–1129.
[10] J.M. Levin, S. Pascarella, P. Argos, J. Garnier, Quantification of secondary structure prediction improvement using multiple alignments, Protein Eng. 6 (1993) 849–854.
[11] H. Wako, T.L. Blundell, Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins, II. Secondary structures, J. Mol. Biol. 238 (1994) 693–708.
[12] C. Geourjon, G. Deleage, SOPM: a self-optimized method for protein secondary structure prediction, Protein Eng. 7 (1994) 157–164.
[13] B. Rost, C. Sander, Combining evolutionary information and neural networks to predict protein secondary structure, Proteins 19 (1994) 55–72.
[14] M.M. Gromiha, P.K. Ponnuswamy, Prediction of protein secondary structures from their hydrophobic characteristics, Int. J. Pept. Protein Res. 45 (1995) 225–240.

[15] W. Kabsch, C. Sander, How good are predictions of protein secondary structure? FEBS Lett. 155 (1983) 179–182.
[16] J. Kyngas, J. Valjakka, Unreliability of the Chou–Fasman parameters in predicting protein secondary structure, Protein Eng. 11 (1998) 345–348.
[17] S.Y. Pasta, B. Raman, T. Ramakrishna, Ch.M. Rao, Role of the conserved SRLFDQFFG region of alpha-crystallin, a small heat shock protein. Effect on oligomeric size, subunit exchange, and chaperone-like activity, J. Biol. Chem. 278 (2003) 51159–51166.
[18] N. Eswar, C. Ramakrishnan, N. Srinivasan, Stranded in isolation: structural role of isolated extended strands in proteins, Protein Eng. 16 (2003) 331–339.
[19] N.K. Dakappagari, J. Pyles, R. Parihar, W.E. Carson, D.C. Young, P.T.P. Kaumaya, A chimeric multi-human epidermal growth factor receptor-2 B cell epitope peptide vaccine mediates superior antitumor responses, J. Immunol. 170 (2003) 4242–4253.
[20] K. Koscielska-Kasprzak, J. Otlewski, Amyloid-forming peptides selected proteolytically from phage display library, Protein Sci. 12 (2003) 1675–1685.
[21] J.P. Malone, A. George, A. Veis, Type I collagen N-telopeptides adopt an ordered structure when docked to their helix receptor during fibrillogenesis, Proteins 54 (2004) 206–215.
[22] D.F. Kelly, K.A. Taylor, Identification of the beta-1 integrin binding site on alpha-actinin by cryoelectron microscopy, J. Struct. Biol. 149 (2005) 290–302.
[23] T.C. Gamblin, Potential structure/function relationships of predicted secondary structural elements of tau, Biochim. Biophys. Acta 1739 (2005) 140–149.
[24] Y. Yin, D. Vafeados, Y. Tao, S. Yoshida, T. Asami, J. Chory, A new class of transcription factors mediates brassino steroid-regulated gene expression in *Arabidopsis*, Cell 120 (2005) 249–259.
[25] H. Seligmann, Cost-minimization of amino acid usage, J. Mol. Evol. 56 (2003) 151–161.
[26] F. Fischel-Ghodsian, G. Mathiowitz, T.F. Smith, Alignment of protein sequences using secondary structure: a modified dynamic programming method, Protein Eng. 3 (1990) 577–581.
[27] M. Murakami, Occurrence of beta-turn potentials around nuclear and nuclear localization sequences, J. Protein Chem. 10 (1991) 469–473.
[28] M. Murakami, Critical amino acids responsible for converting specificities of proteins and for enhancing enzyme evolution are located around beta-turn potentials: data-based prediction, J. Protein Chem. 12 (1993) 783–789.
[29] M. Murakami, Critical amino acids responsible for conferring calcium channel characteristics are located on the surface and around beta-turn potentials of channel proteins, J. Protein Chem. 14 (1995) 111–114.
[30] R. Maclin, J.W. Shavlik, Using knowledge-based neural networks to improve algorithms: refining the Chou–Fasman algorithm for protein folding, Mach. Learn. 11 (1993) 195–215.
[31] K. Park, G.C. Flynn, J.E. Rothman, G.D. Fasman, Conformational change of chaperone Hsc70 upon binding to a decapeptide: a circular dichroism study, Protein Sci. 2 (1993) 325–330.
[32] R. Hiramatsu, M. Abe, M. Morita, S. Noguchi, T. Suzuki, Generalized resistance to thyroid hormone: identification of a novel c-erbA beta thyroid hormone receptor variant (Leu450) in a Japanese family and analysis of its secondary structure by the Chou and Fasman method, Jpn. J. Hum. Genet. 39 (1994) 365–377.
[33] A.A. Salamov, V.V. Solovyev, Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments, J. Mol. Biol. 247 (1995) 11–15.
[34] C.J. Crasto, J. Feng, Sequence codes for extended conformation: a neighbour-dependent sequence analysis of loops in proteins, Proteins 42 (2001) 399–413.
[35] H. Kaur, G.P. Raghava, BTEVAL: a server for evaluation of beta-turn prediction methods, J. Bioinform. Comput. Biol. 1 (2003) 495–504.
[36] J. Wang, J.A. Feng, Exploring the sequence patterns in the alpha-helices of proteins, Protein Eng. 16 (2003) 799–807.
[37] R. Linding, R.B. Russell, V. Neduva, T.J. Gibson, GlobPlot: exploring protein sequences for globularity and disorder, Nucleic Acids Res. 31 (2003) 3701–3708.
[38] J.M. Ball, C.L. Swaggerty, X. Pei, W.S. Lim, X. Xu, V.C. Cox, S.L. Payne, SU proteins from virulent and a virulent EIAV demonstrate distinct biological properties, Virology 333 (2005) 132–144.
[39] R. Xu, Y. Xiao, A common sequence-associated physicochemical feature for proteins of beta-trefoil family, Comput. Biol. Chem. 29 (2005) 79–82.
[40] G. Kugler, R.G. Weiss, B.E. Flucher, M. Grabner, Structural requirements of the dihydropyridine receptor α_{1S} II-III loop for skeletal-type excitation-contraction coupling, J. Biol. Chem. 279 (2004) 4721–4728.
[41] J. Wang, J.A. Feng, NdPASA: a novel pairwise protein sequence alignment algorithm that incorporates neighbor-dependent amino acid propensities, Proteins 58 (2005) 628–637.
[42] P.F. Fuchs, A.J. Alix, High accuracy prediction of beta-turns and their types using propensities and multiple alignments, Proteins 59 (2005) 828–839.
[43] P.Y. Chou, G.D. Fasman, Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins, Biochemistry 13 (1974) 211–222.
[44] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552–557.
[45] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99 (1986) 153–162.
[46] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins 21 (1995) 319–344.
[47] P. Klein, C. DeLisi, Prediction of protein structural class from amino acid sequence, Biopolymers 25 (1986) 1659–1672.
[48] U. Hobohm, M. Scharf, R. Schneider, C. Sander, Selection of a representative set of structures from the Brookhaven Protein Data Bank, Protein Sci. 1 (1992) 409–417.
[49] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[50] B. Rost, V.A. Eyrich, EVA: large-scale analysis of secondary structure prediction, Proteins Suppl. 5 (2001) 192–199.
[51] F. Jiang, Prediction of protein secondary structure with a reliability score estimated by local sequence clustering, Protein Eng. 16 (2003) 651–657.
[52] K.C. Chou, W.M. Liu, G.M. Maggiora, C.T. Zhang, Prediction and classification of domain structural classes, Proteins 31 (1998) 97–103.
[53] Q. Cui, T. Jiang, B. Liu, S. Ma, Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms, BMC Bioinformatics 5 (2004) 66.
[54] B.A. Jameson, H. Wolf, The antigenic index: a novel algorithm for predicting antigenic determinants, Comput. Appl. Biosci. 4 (1988) 181–186.
[55] P.Y. Chou, G.D. Fasman, Prediction of the secondary structure of proteins from their amino acid sequence, Adv. Enzymol. 47 (1978) 45–148.
[56] P.Y. Chou, G.D. Fasman, Empirical predictions of protein conformation, Ann. Rev. Biochem. 47 (1978) 251–276.
[57] K. Nishikawa, Assessment of secondary-structure prediction of proteins. Comparison of computerized Chou–Fasman method with others, Biochim. Biophys. Acta 748 (1983) 285–299.
[58] G.H. Cohen, B. Dietzschold, M. Ponce de Leon, D. Long, E. Golub, A. Varrichio, L. Pereira, R.J. Eisenberg, Localization and synthesis of an antigenic determinant of herpes simplex virus glycoprotein D that stimulates the production of neutralizing antibody, J. Virol. 49 (1984) 102–108.
[59] B.R. Starcich, B.H. Hahn, G.M. Shaw, P.D. McNeely, S. Modrow, H. Wolf, E.S. Parks, W.P. Parks, S.F. Josephs, R.C. Gallo, Identification and characterization of conserved and variable regions

in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS, Cell 45 (1986) 637–648.
[60] M. Motz, J. Fan, R. Seibl, W. Jilg, H. Wolf, Expression of the Epstein-Barr virus 138-kDa early protein in *Escherichia coli* for the use as antigen in diagnostic tests, Gene 42 (1986) 303–312.
[61] S. Modrow, H. Wolf, Characterization of two related Epstein-Barr virus-encoded membrane proteins that are differentially expressed in Burkitt lymphoma and in vitro-transformed cell lines, Proc. Natl. Acad. Sci. USA 83 (1986) 5703–5707.
[62] M.M. Gromiha, S. Selvaraj, Inter-residue interactions in protein folding and stability, Prog. Biophys. Mol. Biol. 86 (2004) 235–277.
[63] M.M. Gromiha, S. Selvaraj, Protein secondary structure prediction in different structural classes, Protein Eng. 11 (1998) 249–251.
[64] F. Eisenhaber, C. Frommel, P. Argos, Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class, Proteins 25 (1996) 169–179.