A new quantitative criterion to distinguish between α/β and α+β proteins (domains)

Chun-Ting Zhang[a,*], Ren Zhang[b]

[a] Department of Physics, Tianjin University, Tianjin 300072, China
[b] Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, China

Received 16 July 1998; received in revised form 19 October 1998

Abstract **According to the statistical analysis, it is shown that the differences of the content of α-helix and β-strand between α/β and α+β proteins are of statistical significance. Based on the secondary structure content and the percentage of parallel or anti-parallel strands, any mixed αβ protein can be represented by a point in a three-dimensional prism. The distribution of the mapping points for 79 mixed αβ proteins (domains), of which 26 are class α/β and 53 are class α+β, shows that the two kinds of points are situated at distinct regions roughly. A new quantitative criterion based on the Fisher discriminant algorithm is proposed to distinguish between the α/β and α+β proteins (domains). Of the 79 proteins 77 are correctly classified (97.5%). As a stringent cross-validation test, the jackknife test shows that of the 79 proteins 77 are correctly classified. The jackknife test accuracy is still 97.5%. These figures indicate the self-consistence and the extrapolating effectiveness of the new quantitative criterion. Applying the new criterion to reclassify the α/β and α+β proteins (domains) in SCOP is also discussed. It is hoped that the new quantitative criterion will be useful for the development of protein classification databases.**

Key words: α/β protein; α+β protein; Secondary structure content; Percentage of parallel strands; Quantitative criterion; Fisher discriminant algorithm

1. Introduction

The concept of protein structural class was first proposed by Levitt and Chothia in 1976 [1]. According to this concept, a globular protein can be assigned to one of the four structural classes, i.e. all-α, all-β, α+β and α/β. The all-α and all-β proteins were defined to be composed of almost entirely α-helices and β-strands, respectively. The α+β proteins were defined to be composed of separate segments of α-helices and β-strands (mainly anti-parallel), whereas the α/β proteins were defined to be composed of mixed segments of α-helices and β-strands (mainly parallel). Because there were very few proteins whose crystallographic structures were known in 1976, this definition of the structural classes was derived from a quite small database, i.e. 31 globular proteins only [1]. Now the three-dimensional structures of about 6000 proteins are known. However, the definition of the protein structural classes of Levitt and Chothia [1] is still accepted by the protein research community even to date. Since 1976, many definitions of the structural classes have been proposed along with the work of Levitt and Chothia [1], for example, Nakashima et al. [2], P.Y. Chou [3], Scheridan et al. [4], Klein and DeLisi [5], Kneller et al. [6], K.-C. Chou [7], and Michie et al. [8]. In these studies, the classification schemes are basically based on the secondary structure content of proteins. Only few researchers provide quantitative criteria to distinguish between the α/β and α+β proteins [7,8]. K.-C. Chou demands that the mixed αβ proteins (domains) should be classified as α/β if the percentage of parallel strands is greater than 60%, otherwise, if the percentage of anti-parallel strands is greater than 60%, it should be an α+β protein [7]. On the other hand, to separate the α/β and α+β proteins, Michie et al. introduced a new parameter called the alternation score of secondary structures along the polypeptide chain [8]. The mixed αβ protein (domain) is mapped onto a point in the two-dimensional plane spanned by the alternation score and percentage of parallel strands. Based on the distribution of the mapping points, a quantitative criterion was proposed to classify the mixed αβ proteins (domains) into the α/β and α+β classes [8]. Obviously, the threshold of 60% adopted by K.-C. Chou [7] is a simple majority only. The separation between the α/β and α+β classes based on the criterion of Michie et al. [8] is more objective and hence more reliable. However, in the work of Michie et al. [8], the role played by the secondary structure content in the separation of the α/β and α+β classes was almost ignored. Based on the proteins in the training set (see below), we have performed a Student's *t*-test to examine the null hypothesis H_0: there is no difference of the content of α-helix and β-strand between the α/β and α+β classes. Consequently, the hypothesis is rejected. In other words, the differences of the content of α-helix and β-strand between the α/β and α+β classes are of statistical significance. Based on this analysis, a new quantitative criterion to distinguish between the α/β and α+β classes is proposed here, using the secondary structure content and the percentage of the parallel strands as well. As we will see later, high classification accuracy has been achieved using the new quantitative criterion.

2. Materials and methods

Recently, 210 representative non-homologous proteins (domains) were classified manually by Michie et al. [8,9]. Of the 210 proteins, 56 proteins are classified as all-α, 75 all-β, 26 α/β and 53 α+β. In this study, the 26 α/β and 53 α+β proteins (domains) are used as the training set. From the Protein Data Bank, all the three-dimensional structures of the 79 proteins stored as PDB files can be obtained. Their PDB codes are listed in Table 1.

To test the quantitative criterion derived from the proteins in the training set, we need some other proteins as a test set, which should be independent of the training set. There were another four α/β and 11 α+β proteins (domains) which were also classified manually [8]. It was shown that some ambiguity occurred when these proteins were classified by an automated class assignment protocol proposed by Michie et al. [8]. It was thought that these proteins (domains) are in the borderline region and left for manual inspection only [8]. In order to test the new quantitative criterion more stringently, these 15 pro-

*Corresponding author. Fax: (86) (22) 23358329.
E-mail: ctzhang@tju.edu.cn

PII: S0014-5793(98)01433-1

Table 1
The PDB codes of the proteins (domains) in the training set

α/β	1xis_	5timA	1nar_	2mnr_	1chrA	1fbaA	1gox_	5p21_
	3chy_	1etu_	1ofv_	4fxn_	1cseE	1cde_	2trxA	3trx_
	1gp1A	4dfrA	2ak3A	3adk_	2ctc_	2cmd_	1ipd_	7icd_
	1nipA	1tml_						
α+β	1gps_	2ovo_	1tgsI	1gatA	1ptf_	2bopA	1cewI	1stfI
	1ctf_	1fxd_	2nckL	3rubS	1pba_	1aps_	1cseI	2sicI
	2rn2_	1aak_	7rsa_	1onc_	1fus_	1brnL	1pgx_	1frrA
	1ubq_	2sns_	1ltsD	3il8_	1fkb_	2msbA	5pti_	1and_
	1zaaC	1shaA	2pna_	1poc_	3cla_	1eaf_	1ltsA	2tscA
	1rveA	3b5c_	1cmbA	2cpl_	5fd1_	2dnjA	1mat_	1pyaB
	1hgeB	1pkp_	3monB	1bw3_	1vil_			

teins are used as the test set in this study. They are thought of as a touchstone to test the quantitative criterion proposed here to distinguish between the α/β and α+β proteins (domains). The PDB codes of these 15 proteins (domains) are listed in Table 2. The secondary structure content of the above proteins (domains) is determined by the DSSP method [10].

Denoted by α, β and c the content of α-helix, β-strand and coil in a protein (domain), respectively, we find

$$\alpha + \beta + c = 1. \tag{1}$$

This means that of the three real numbers α, β and c only two are independent. Note that there is a simple theorem about a regular triangle with its height equal to 1. The sum of the three distances to the three sides of the regular triangle of any point within it equals exactly 1. Consequently, the three real numbers α, β and c can be represented by a point in this triangle. Set up an appropriate coordinate system such that the origin coincides with the center of the triangle and the x-axis is parallel with one of the three sides. The coordinate x and y associated with the three numbers can be expressed in terms of α, β and c. Simple geometrical calculation shows that

$$\begin{cases} x = (\beta - \alpha)/\sqrt{3}, \\ y = 2/3 - (\alpha + \beta), \end{cases} \tag{2}$$

Meanwhile, the percentage of the parallel strands in a protein (domain), denoted by z, is calculated from the output file of the DSSP program, and the counting unit is the amino acid, not the strand, as described in detail by Chou [7]. Consequently, each protein can be represented by a mapping point or a vector in a three-dimensional (3D) space, spanned by x, y and z. Obviously, the actual shape of the space in which the mapping points are distributed is a 3D prism. Denoting the vector representing the ith protein in the 3D prism by $\mathbf{r}_i$, we have

$$\mathbf{r}_i = (x_i,\ y_i\ z_i)^T,\ i = 1,\ 2,\ ...,\ N, \tag{3}$$

where T is a transpose operation for a matrix and

$$\begin{cases} x_i = (\beta_i - \alpha_i)/\sqrt{3}, \\ y_i = 2/3 - (\alpha_i + \beta_i), \end{cases} \tag{4}$$

where α_i, β_i and z_i are the content of α-helix, β-strand and the percentage of the parallel strands, respectively, in the ith protein in the training set with N proteins.

The Fisher linear discriminant algorithm [11] is used here to find an appropriate plane in the 3D prism to distinguish between the two kinds of mapping points, one represents the α/β and another α+β proteins. This plane is described by the following equation

$$c_1 x + c_2 y + c_3 z = t, \tag{5}$$

where c_1, c_2 and c_3 are three parameters describing the plane and t is an appropriate threshold. The vector $\mathbf{c}$ with the three components c_1, c_2 and c_3 is used to represent the three parameters. Both $\mathbf{c}$ and t are determined by the data $\mathbf{r}_i$, $i = 1, 2, ..., N$, in the training database. The procedure to determine $\mathbf{c}$ is described in any book of multi-linear analysis, e.g. Mardia et al. [11]. The vector $\mathbf{c}$ is not unique in the sense that $\mathbf{c}$ multiplied by a constant is still acceptable. Without losing generality we choose the constant such that $|\mathbf{c}|^2 = 1$. The threshold t is determined by the requirement that the percentage accuracy for distinguish between the α/β or α+β proteins reaches the maximum. The percentage accuracy is defined as the fraction of proteins in the training set, which are correctly discriminated by the Fisher plane described by Eq. 5. This percentage accuracy is used to test the self-consistency of the discriminant algorithm. So, it is also called the accuracy of re-substitution. Once the vector $\mathbf{c}$ and the threshold t are obtained, the decision of α/β or α+β for each protein in the test set is simply performed by the criterion of $\mathbf{c}\cdot\mathbf{r} > t$ or $\mathbf{c}\cdot\mathbf{r} < t$, where $\mathbf{r} = (x, y, z)^T$, where x and y are defined by Eq. 2, and z is the percentage of parallel strands in the protein concerned. Two-fold cross-validation tests are performed to evaluate the new quantitative criterion. One is the jackknife analysis and another is the single-test-set analysis. The evaluation of the quantitative discriminant criterion is simply determined by the percentage accuracy, a fraction of proteins in the test set, which are correctly discriminated by the criterion of $\mathbf{c}\cdot\mathbf{m} > t$ or $\mathbf{c}\cdot\mathbf{m} < t$. This percentage accuracy is used to test the extrapolating effectiveness of the discriminant algorithm. So, it is also called the test accuracy.

3. Result and discussion

Based on the data derived from the training set, the vector $\mathbf{c}$ and the appropriate threshold t are determined. The decision of α/β class or α+β class for each protein in the training set is simply performed by the criterion of $\mathbf{c}\cdot\mathbf{r} > t$ or $\mathbf{c}\cdot\mathbf{r} < t$, where $\mathbf{r} = (x, y, z)^T$, where x and y are defined by Eq. 2, and z is the percentage of parallel strands in the protein concerned. Consequently, of the 26 α/β proteins (domains) in the training set, 26 are correctly classified. Of the 53 α+β proteins (domains) in the training set, 51 are correctly classified. On average, the accuracy of re-substitution is (26+51)/(26+53) = 77/79 = 97.5%, indicating a high self-consistency of the quantitative criterion. Only the α+β proteins (domains) 1cseI and 2rn2_ are incorrectly classified as α/β proteins (domains). The content of helix, strand and the percentage of parallel strands for the former are 0.17, 0.30 and 0.55, respectively; and for the latter are 0.35, 0.28 and 0.38, respectively. As we can see from the above figures, the mapping points of both proteins are in the borderline region between the α/β class and α+β class. Their

Table 2
The PDB codes of the proteins (domains) in the test set

α/β	1aba	1ego	3dfr	8dfr				
α+β	1bovA	1cbn	1fdx	1gmpA	1hev	1hrhA	1poa	1pyp
	2sn3	9wgaA	8rxnA					

Table 3
Means and standard deviations of helix/strand content for the *t*-test[a]

Content	Class	N[b]	Mean	S.D.	P value[c]
α-helix	α/β	26	0.375	0.074	
	α+β	53	0.235	0.090	<0.001
β-strand	α/β	26	0.183	0.057	
	α+β	53	0.260	0.087	<0.001

[a]The means and standard deviation of the helix content for the 26 α/β and 53 α+β proteins (domains) are listed in the first and second row, respectively. Those of the strand content are in the third and fourth row, respectively.
[b]Number of proteins in the training set.
[c]P denotes the probability that the difference of the means of content between the α/β and α+β proteins (domains) is caused by random events. Since $P<0.001$, the null hypothesis H_0 is not valid.

structural classes cannot be assigned correctly by the Fisher plane in the three-dimensional prism in this case.

To test the extrapolating effectiveness of the new quantitative criterion, a stringent cross-validation test, jackknife analysis, is performed, which is deemed to be one of the most effective and objective cross-validation tests. The jackknife test is also called the leave-one-out test (see, e.g. Mardia et al. [11]), in which each protein (domain) is in turn singled out as a tested sample and the vector **c** and the threshold t are derived without using this protein (domain). In other words, the classification of the singled out protein (domain) is performed by the vector **c** and the threshold t derived using all other proteins (domains) except the one that is being classified. In each jackknife stage, the singled out protein (domain) is classified by the criterion of $\mathbf{c}\cdot\mathbf{r}>t$ or $\mathbf{c}\cdot\mathbf{r}<t$. Consequently, of the 79 proteins (domains) 77 are correctly classified. The jackknife test accuracy is 77/79 = 97.5%, indicating a high extrapolating effectiveness of the new quantitative criterion. To test the new quantitative criterion further, the cross-validation using a single test set is also performed. As mentioned previously that there were another four α/β and 11 α+β proteins (domains) which were also classified manually [8]. Interestingly, of the four α/β proteins in the test set, three are correctly classified. Only the α/β protein 1ego_ is incorrectly classified as an α+β protein. Of the 11 α+β proteins (domains) in the same test set, all 11 are correctly classified. On average, the test accuracy is 14/15 = 93.3%, also indicating a high extrapolating effectiveness of the quantitative criterion.

The above results indicate that the new quantitative criterion to distinguish between the α/β and α+β proteins (domains) is not only very simple, but also very effective. The high accuracy obtained for distinguishing between the α/β and α+β proteins (domains) also implies that the classification of the α/β and α+β proteins (domains) is basically determined by the secondary structure content and the percentage of parallel or anti-parallel strands. The percentage of parallel or anti-parallel strands plays an important role in distinguishing between the two structural classes [7,8]. However, our study shows that, in addition to the percentage of parallel or anti-parallel strands, the secondary structure content also plays a key role. We have performed a *t*-test to see if the differences of the secondary structure content between the α/β and α+β proteins have statistical significance. The *t*-test is based on the 79 proteins in the training set. The null hypothesis H_0 is: there is no difference of the population of the secondary structure content between the α/β and α+β proteins. The 26

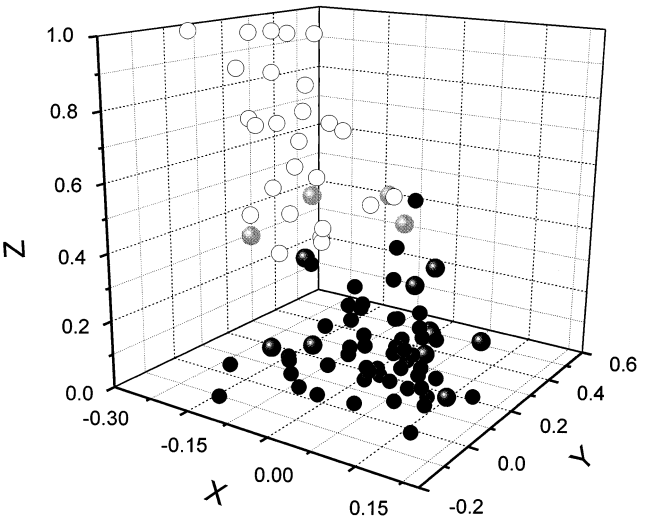

Fig. 1. The distribution of the mapping points in the three-dimensional space spanned by x, y and z, where x and y are defined by Eq. 2 and z is the percentage of parallel strands. The mapping points representing the α/β proteins in the training set and the test set are denoted by white circles and gray spheres, respectively. The mapping points representing the α+β proteins in the training set are denoted by black circles, and those in the test set are denoted by black spheres. Note that the two kinds of mapping points are situated at distinct regions, which constitutes the basis to distinguish between them.

α/β and 53 α+β proteins in the training set are regarded as samples from the respective population. The means and standard deviations of the content of helix/strand for the 26 α/β and 53 α+β proteins are listed in Table 3. The result of the *t*-test shows that the null hypothesis H_0 is not valid for both content of α-helix and β-strand. In other words, the null hypothesis H_0 is actually rejected. The differences of the secondary structure content between the two classes are really of statistical significance. The distribution of the two kinds of mapping points in the 3D prism is shown in Fig. 1, where those representing the α/β proteins and α+β proteins in the training set are denoted by white and black circles, respectively. The mapping points representing the α/β and α+β proteins in the test set are denoted by gray and black spheres, respectively. As we can see, the two kinds of mapping points are distributed roughly in two distinct regions. This specific distribution constitutes the basis to distinguish between the α/β and α+β proteins (domains) in this study.

Compared with other criteria, we find that of the 79 proteins in the training set only 69 are correctly classified according to Chou's criterion [7]. The percentage accuracy is 69/79 = 87.3%. Of the 15 proteins in the test set only 11 are

Table 4
The vector **c** and threshold t[a]

c_1	c_2	c_3	t
0.0152	−0.2517	0.9677	0.3150

[a]The Fisher discriminant parameters $\mathbf{c}=(c_1,c_2,c_3)$ and the threshold t are calculated based on the 94 proteins (domains) including those in the training set and test set. These parameters are provided for users to classify the protein (domain) into the α/β or α+β class. The decision of the α/β class or α+β class is performed by the criterion $\mathbf{c}\cdot\mathbf{r}>t$ or $\mathbf{c}\cdot\mathbf{r}<t$, where $\mathbf{r}=(x,y,z)^{\mathrm{T}}$, x and y are defined by Eq. 2 and z is the percentage of parallel strands.

Table 5
The description of the 125 mixed αβ proteins (domains) in SCOP[a]

	PDB code	region	PDB code	region	PDB code	region
α/β	1cgt_	| 1–382	1cxe_	| 1-382	1cgv_	| 1-382
	1btb_	| W.C.	1brsD	| W.C.	1cxf_	| 1-382
	1fnd_	| 155–314	4ts1A	| 1–217	1selA	| W.C.
	1cdoA	| 176–324	1hldA	| 175–324	1horA	| W.C.
	2secE	| W.C.	1cia_	| W.C.	1frn_	| 155–314
	1pnt_	| W.C.	2hnp_	| W.C.	1tybE	| 1–217
	1tho_	| W.C.	1tkbA	| 535–680	1lam_	| 1–159
	1bllE	| 1–159	1gdtA	| 1–140	3hsc_	| 3–188
	1idm_	| W.C.	1ngi_	| 4–188	1atr_	| 2–188
	1cde_	| W.C.	1grcA	| W.C.	1cddA	| W.C.
	1mhtA	| W.C.	1ama_	| W.C.	1alhA	| W.C.
	1ula_	| W.C.	1ngb_	| 4–188	1rhd_	| 1–149
	1trx_	| W.C.	1amn_	| W.C.	8atcA	| 1–150
	1acj_	| W.C.	1alkA	| W.C.	2ctc_	| W.C.
	1dr1_	| W.C.	1drj_	| W.C.	1hqaA	| W.C.
	1ajdA	| W.C.	1acl_	| W.C.	1ngg_	| 3–188
	1ajcA	| W.C.	1dbp_	| W.C.	1xab_	| W.C.
	1raiA	| 1–150	1scnE	| W.C.	1ttqB	| W.C.
	1wsyB	| W.C.	1orb_	| 1–149	1ajaA	| W.C.
	2anhA	| W.C.	5acn_	| 1–528	5cpa_	| W.C.
	2bgt_	| W.C.	1drk_	| W.C.	1acmA	| 1–150
	1ngh_	| 4–188	1olcA	| W.C.	1ctu_	| 1–150
α+β	1fut_	| W.C.	2baa_	| W.C.	1aec_	| W.C.
	2rat_	| W.C.	2rns_	| W.C.	1ras_	| W.C.
	1ssbA	| W.C.	1rbd_	| W.C.	1kraA	| W.C.
	1pgx_	| W.C.	1pgb_	| W.C.	1igcA	| W.C.
	2igg_	| W.C.	2igh_	| W.C.	2secI	| W.C.
	1coy_	| 319–450	3monA	| W.C.	1frtA	| 1–178
	1fkj_	| W.C.	2tecI	| W.C.	1lttA	| W.C.
	1egl_	| W.C.	1sbnI	| W.C.	3mdsA	| 93–203
	1vig_	| W.C.	1egpA	| W.C.	1fkl_	| W.C.
	1mns_	| 3–132	1grl_	| 137–190	1fccC	| W.C.
	1rldS	| W.C.	1comA	| W.C.	1sphA	| W.C.
	1gaeO	| 149–312	1mstA	| W.C.	1grb_	| 364–478
	1lklA	| W.C.	1lcjA	| W.C.	1lckA	| 117–226
	1sceA	| W.C.	1setA	| 111–421	1sibI	| W.C.
	1tsdA	| W.C.	1htlA	| W.C.	1bmsA	| W.C.
	2hpr_	| W.C.	1tsy_	| W.C.	1tys_	| W.C.
	3b5c_	| W.C.	1tbpA	| 61–155	1xrc_	| 1–101
	1glv_	| 123–316	2tscA	| W.C.	3dni_	| W.C.
	1dnkA	| W.C.	4mdhA	| 155–333	1mrk_	| W.C.
	1ltaA	| W.C.	1ltgA	| W.C.		

[a]Each protein (domain) is expressed by a symbol A|B, where A is the corresponding PDB code, and B is the sequence region. When a domain is constituted by whole chain, B = W.C.; otherwise, B contains two numbers to indicate its starting and end points along the sequence. The classification of these 125 proteins (domains) is based on SCOP [12].

correctly classified using Chou's criterion. The percentage accuracy is 11/15 = 73.3% only. According to Michie et al., all 15 proteins in the test set cannot be classified by the automated class assignment protocol proposed by them due to the borderline effect [8]. Compared with the above two criteria, the advantage of our criterion is obvious. First of all, the new criterion has a high accuracy for both the re-substitution test and the jackknife test. Second, for those proteins (domains) in the borderline region, our criterion is capable of classifying them with a high accuracy (93.3% is found in this study). The new criterion proposed here to distinguish between the α/β and α+β proteins (domains) is at least a complement to the classification criteria currently available.

After we finished the above analysis, the 94 proteins (domains) including those in the training set and test set were merged together, forming a larger new training set. In this set, there were 30 α/β and 64 α+β proteins (domains). The Fisher discriminant algorithm was also applied to this set. The corresponding vector **c** and the threshold t were obtained, as listed in Table 4. These parameters will be used to distinguish between the α/β and α+β proteins (domains) for the users. The decision is simply performed by the criterion of $\mathbf{c}\cdot\mathbf{r} > t$ or $\mathbf{c}\cdot\mathbf{r} < t$. More clearly, for a given protein (domain) to be classified, calculate $\mathbf{c}\cdot\mathbf{r} = c_1x + c_2y + c_3z$. If it is greater than t, the protein studied is classified as an α/β protein (domain), otherwise if $\mathbf{c}\cdot\mathbf{r} < t$, it is classified as an α+β protein (domain).

It is very interesting to apply the above new criterion to the SCOP database [12]. Recently, 125 mixed αβ proteins (domains) in SCOP were used for predicting the structural classes, of which there were 66 α/β and 59 α+β proteins (domains) [13]. Although these mixed αβ proteins (domains) are by no means all of those classified as α/β and α+β in SCOP at present, they are representatives of the relevant protein families and super-families. The PDB codes of these 125 proteins (domains) are listed in Table 5. Applying the new criterion and the parameters listed in Table 4 to reclassify the 125 proteins (domains), we find that of the 66 α/β and 59 α+β proteins (domains) 55 and 51, respectively, are correctly re-

Table 6
The parameters of the 19 proteins (domains) in SCOP incorrectly reclassified[a]

α/β				α+β			
PDB code	α %	β %	z	PDB code	α %	β %	z
1cia_	0.28	0.29	0.21	2secI	0.17	0.30	0.55
2hnp_	0.31	0.21	0.21	2tecI	0.17	0.25	0.66
3hsc_	0.27	0.31	0.25	1egl_	0.16	0.20	0.79
1ngi_	0.29	0.31	0.25	1sbnI	0.17	0.30	0.55
1atr_	0.30	0.30	0.25	1egpA	0.18	0.23	0.75
1mhtA	0.27	0.18	0.34	1grl_	0.46	0.15	0.62
1ngb_	0.28	0.31	0.25	1sibI	0.17	0.29	0.61
1ngg_	0.28	0.31	0.26	1xrc_	0.31	0.39	0.35
1ngh_	0.30	0.31	0.25				
1olcA	0.28	0.20	0.22				
1ctu_	0.36	0.15	0.16				

[a]The left part indicates the 11 α/β proteins (domains) incorrectly reclassified as α+β and the right part indicates the eight α+β proteins (domains) incorrectly reclassified as α/β. The PDB codes, the content of α-helix and β-strand and the percentage of the parallel strands z are listed here.

classified. In other words, 19 are incorrectly reclassified, of which 11 α/β are reclassified as α+β and eight α+β are reclassified as α/β. The resulting classification accuracy is only (125−19)/125 = 84.8%, much lower than 97.5% for classifying the proteins (domains) in the database of Michie et al. [8]. The 125 proteins (domains) were also reclassified by Chou's criterion [7]. Accordingly, the reclassification accuracy was only 65.6%. The remarkable difference between the two figures (84.8% and 97.5%) reflects the different criteria for classifying the mixed αβ proteins (domains) between the two databases. In the database of Michie et al. [8], the parallel and anti-parallel strands are the key factor to distinguish between the α/β and α+β proteins (domains), while in SCOP, the above factor (parallel and anti-parallel) is not always the decision one. To illustrate this, for example, the 19 proteins (domains) incorrectly reclassified by the present method are described in Table 6, in which the content of α-helix, β-strand and the percentage of parallel strands are listed. We can see that in the 11 α/β proteins (domains) in SCOP the β-strands are basically anti-parallel, while in the eight α+β proteins (domains) in SCOP the β-strands are basically parallel. This is completely contrary to the original idea for distinguishing between α/β and α+β of Levitt and Chothia [1]. To solve this contradiction, in addition to the secondary structure content and the percentage of the parallel strands, other parameters should be introduced to distinguish between the two classes with much higher accuracy. Let us consider another example to elucidate the necessity of introducing other classification parameters. A TIM barrel β8α8 consisting of eight parallel strands is obviously an α/β domain and would be very similar to a hypothetical β8α8 barrel possessing an all-anti-parallel β strand. This structure will be classified as α+β by the new classification criterion. However, if the secondary structure alternation score proposed by Michie et al. [8] is used in the classification procedure, the hypothetical β8α8 barrel may be still classified as α/β. Therefore, once a complete set of classification parameters is set up, the classification of α/β and α+β proteins could be solved satisfactorily by using the methodology presented in this paper.

In conclusion, the distinction between the α/β and α+β proteins (domains) is not only possible but also necessary. As shown in this study, the difference of the secondary structure content between the two classes is of statistical significance. This means that the distinction between the two classes is objective, rather than subjective. Furthermore, it will be worthwhile for the structure and function prediction of proteins based on the differentiation of α/β and α+β. This study may be considered a first step towards distinguishing between the α/β and α+β proteins (domains) with a reliable quantitative criterion. It is hoped that our work will be useful for the development of protein classification databases, such as SCOP [12] etc. The relevant computer programs are available on request.

Acknowledgements: We are grateful to Dr. K.-C. Chou for helping us calculate the percentages of the parallel strands for the proteins (domains) concerned in this study. We are also grateful to the anonymous referees for carefully reviewing the manuscript and the constructive suggestions which improved the presentation significantly. The present study was supported in part by the Pandeng Project of China and a grant from the State Education Commission of China.

References

[1] Levitt, M. and Chothia, C. (1976) Nature 261, 552–557.
[2] Nakashima, H., Nishikawa, K. and Ooi, T. (1986) J. Biochem. 99, 153–162.
[3] Chou, P.Y. (1989) in: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D., Ed.), pp. 549–586, Plenum Press, New York.
[4] Sheridan, R.P., Dixon, J.S., Venkataraghavan, R., Kuntz, I.D. and Scott, K.P. (1985) Biopolymers 24, 1995–2023.
[5] Klein, P. and DeLisi, C. (1986) Biopolymers 25, 1659–1672.
[6] Kneller, D.G., Cohen, F.E. and Langridge, R. (1990) J. Mol. Biol. 214, 171–182.
[7] Chou, K.-C. (1995) Proteins 21, 319–344.
[8] Michie, A.D., Orengo, C.A. and Thornton, J.M. (1996) J. Mol. Biol. 262, 168–185.
[9] Michie, A., Hutchinson, E., Laskowski, L., Orengo, C. and Thornton, J. (1995) in: Making the Most of Your Model, Proceeding of the CCP4 Study Weekend (Hunter, W., Thornton, J. and Bailey, S., Eds.), pp. 83–94, Council for the Central Laboratory of the Research Councils.
[10] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[11] Mardia, K.V., Kert, J.H. and Bibby, J.M. (1979) Multivariate Analysis, Academic Press, London.
[12] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.
[13] Chou, K.-C., Liu, W.-M., Maggiora, G.M. and Zhang, C.-T. (1998) Proteins 31, 97–103.