J. theor. Biol. (1993) **161**, 251–262

A Joint Prediction of the Folding Types of 1490 Human Proteins from their Genetic Codons

JAMES JEIWEN CHOU† AND CHUN-TING ZHANG‡§

† *Department of Physics, University of Michigan, Ann Arbor, MI* 48109, *U.S.A. and* ‡ *Department of Physics, Tianjin University, Tianjin* 300072, *China*

(*Received on* 14 *May* 1992, *Accepted in revised form on* 4 *September* 1992)

The codon usages for 1490 human proteins have been published by Wada *et al.* (1990). Based on these data, the frequencies of occurrence of 20 amino acids for each of the 1490 proteins have been calculated according to the genetic codes. Proteins are generally classified into five folding types, i.e. the α, β, $\alpha+\beta$, α/β and ζ (irregular) types. The folding type of a protein is correlated to its amino acid composition. By means of three methods established by different investigators, the folding type for each of the 1490 human proteins has been predicted. It has been demonstrated that the accuracy of prediction for the 1490 human proteins is at least 80% by examining the predicted results of some structurally known proteins with these methods. There are only six proteins for which there is uncertainty about their folding types as completely inconsistent results were obtained when predicted with the three different methods. For the remaining 1484 human proteins the numbers of α, β, $\alpha+\beta$, α/β, and ζ folding type proteins were found to be 128, 235, 169, 933 and 19, respectively, suggesting that the α/β type proteins would predominate in this set of human proteins. The occurrence frequencies of bases in the first, second and third codon position for each folding type of protein have been calculated. It is shown that the folding type of a protein is strongly dependent on the ratio of frequency of base G in the first codon position with that in the second codon position. The biological implication of the results has been discussed.

Introduction

In their pioneering works in 1980 and 1981, Grantham and his colleagues (Grantham *et al.*, 1980, 1981) reported the codon usage in a total of 161 protein genes then available. Since then, the size of the database has grown larger and larger. The codon usage in 1638, 3681 and 11 415 genes were compiled by Ikemura and his colleagues in 1986, 1988 and 1990, respectively (Maruyama *et al.*, 1986; Aota *et al.*, 1988; Wada *et al.*, 1990). The codon usage reported by Wada *et al.* (1990) is the newest and largest set. We regard the codon usage patterns for different proteins in various organisms as "a book written by God". It is thrilling to "read" and analyze such a "book". The genetic codes establish a definite connection between each amino acid with one or several codons (in the case of degeneracy). Therefore, once the codon usage is known, the number of occurrences of each amino acid in a protein can be

§ Author to whom correspondence should be addressed.

0022-5193/93/060251+12 $08.00/0

easily calculated by the genetic codes. In other words, the frequencies of occurrence of amino acids can be easily obtained. So far, the codon usages for 1490 human proteins have been published (Wada *et al.*, 1990). Based on these data and the genetic codes (Nirenberg, 1963; Marshall *et al.*, 1967) given in Table 1, the frequencies of amino acids of these proteins have been calculated in this paper.

TABLE 1

The genetic code: a summary of the triplet codons

First	Second U	C	A	G	Third
U	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	Ter†	Ter†	A
U	Leu	Ser	Ter†	Trp	G
C	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
C	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
A	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
G	Val	Ala	Glu	Gly	G

† Ter means termination.

On the other hand, it is well known that the proteins of known structure may be classified into five folding types, namely, the α, β, $\alpha+\beta$, α/β, and ζ (irregular) types (Levitt *et al.*, 1976; Richardson & Richardson, 1989). It was also found that the folding type of a protein is relevant to its amino acid composition (Chou, 1980; Nakashima *et al.*, 1986). Therefore, the folding type of a protein can be predicted according to its frequencies of occurrence of amino acids. In fact, the prediction of folding types for 64 structurally known proteins based only on their amino acid composition are considerably successful (Chou, 1980). The average prediction accuracy is nearly 80%. Using a different method, Nakashima *et al.* (1986) made a similar prediction for 135 structurally known proteins, further confirming such a correlation between the amino acid composition of a protein and its folding type. Recently, a new prediction has been proposed (Chou & Zhang, 1992*a*). It was found that by means of this new method the rate of correct prediction for the α-type proteins could reach as high as 97·4%, with the average accuracy rate of 83·6% for all folding types of proteins. The present study was initiated by these encouraging results in an attempt to predict the folding types of the 1490 human proteins, based on their amino acid composition derived from the codon usage data. In order to increase the reliability of our prediction, a combination of three different methods, i.e. the method by Chou (1980, 1989), the one by Nakashima *et al.* (1986), and the one by Chou & Zhang (1992*a*), has been adopted as will be detailed below.

In addition, the occurrence frequencies of bases in the first, second and third codon position of the sequences coding for the proteins of each folding type have been calculated and discussed. The mean length of proteins for each of the five folding types has been calculated as well.

Method

Three different prediction methods will be briefly described here. To elaborate the methods in a unified manner, it is convenient to use the concept of composition space first introduced by Nakashima *et al.* (1986). An Euclidean space spanned by 20 frequencies of amino acid is defined as the composition space. It is a 20-dimensional space. For any protein, there are 20 frequencies of amino acid. Therefore, each protein can be represented by a definite point or a vector in the composition space. The principle of these methods is that the folding type of a protein is correlated to its amino acid composition. According to Nakashima *et al.* (1986), the definitions of the five protein folding types can be quantitatively described as follows. (i) α type: proteins of this type contain more than 15% α-helices and less than 10% β-strands; (ii) β type: proteins of this type contain less than 15% α-helices and more than 10% β-strands; (iii) $\alpha+\beta$ type: proteins of this type contain more than 15% α-helices and more than 10% β-strands with dominantly antiparallel β-strands; (iv) α/β type: proteins of this type contain more than 15% α-helices and more than 10% β-strands with dominantly parallel β-strands; (v) ζ type: proteins of this type contain less than 15% α-helices and less than 10% β-strands. Based on 135 structurally known proteins, the standard amino acid compositions for the five protein folding types have already been derived by Nakashima *et al.* (1986) as shown in Table 2. The five standard amino acid compositions listed in five columns of this table are denoted hereafter by five vectors $\boldsymbol{v}(\alpha)$, $\boldsymbol{v}(\beta)$, $\boldsymbol{v}(\alpha+\beta)$, $\boldsymbol{v}(\alpha/\beta)$, and $\boldsymbol{v}(\zeta)$, representing the norms of α, β, $\alpha+\beta$, α/β, and ζ proteins, respectively. We will use these norms to predict the folding types of the 1490 human proteins.

METHOD A (CHOU & ZHANG, 1992*a*)

Suppose that $\boldsymbol{v}(x)=\{v_1(x), v_2(x), \ldots, v_{20}(x)\}$ is a vector in the composition space, where $v_i(x)$ represents the relative frequency of amino acid i occurring in the protein x to be predicted. The dot product of $\boldsymbol{v}(x)$ with each of the five standard vectors $\boldsymbol{v}(k)$ is calculated, respectively, as follows:

$$\boldsymbol{v}(x) \cdot \boldsymbol{v}(k)=\sum_{i=1}^{20} v_i(x)v_i(k) \qquad (k=\alpha, \beta, \alpha+\beta, \alpha/\beta, \zeta). \tag{1}$$

A correlation angle $\theta(k)$ between $\boldsymbol{v}(x)$ and $\boldsymbol{v}(k)$ is defined via

$$\boldsymbol{v}(x) \cdot \boldsymbol{v}(k)=|\boldsymbol{v}(x)||\boldsymbol{v}(k)| \cos \theta(k). \tag{2}$$

TABLE 2

The five standard vectors representing the norms of the five folding types in the 20-D composition space†

Amino acid Order‡	Name	Standard vector $v(\alpha)$	$v(\beta)$	$v(\alpha+\beta)$	$v(\alpha/\beta)$	$v(\zeta)$
1	Arg	0·0279	0·0322	0·0405	0·0435	0·0108
2	Leu	0·0889	0·0669	0·0637	0·0854	0·0402
3	Ser	0·0544	0·0950	0·0705	0·0589	0·0642
4	Thr	0·0491	0·0783	0·0641	0·0550	0·0435
5	Pro	0·0381	0·0523	0·0429	0·0436	0·0582
6	Ala	0·1163	0·0754	0·0889	0·0883	0·0890
7	Gly	0·0766	0·0987	0·0800	0·0871	0·1049
8	Val	0·0602	0·0748	0·0650	0·0762	0·0489
9	Lys	0·1010	0·0466	0·0718	0·0655	0·0327
10	Asn	0·0379	0·0490	0·0560	0·0413	0·0416
11	Gln	0·0333	0·0412	0·0317	0·0344	0·0403
12	His	0·0279	0·0164	0·0200	0·0219	0·0102
13	Glu	0·0652	0·0375	0·0618	0·0612	0·0685
14	Asp	0·0652	0·0537	0·0576	0·0612	0·0885
15	Tyr	0·0255	0·0367	0·0459	0·0302	0·0395
16	Cys	0·0171	0·0348	0·0294	0·0143	0·1204
17	Phe	0·0422	0·0357	0·0360	0·0388	0·0173
18	Ile	0·0372	0·0476	0·0474	0·0582	0·0699
19	Met	0·0242	0·0124	0·0140	0·0214	0·0053
20	Trp	0·0117	0·0148	0·0128	0·0138	0·0062

† The five standard vectors represent all-α, all-β, $\alpha+\beta$, α/β, and irregular (ζ) type proteins, respectively. Each such vector has 20 components, which are defined by the average amino acid compositions of the corresponding folding type, and whose values are derived from the 135 structurally known proteins (Nakashima *et al.*, 1986).

‡ The order of amino acids, each of which corresponds to a component of the 20-dimensional composition space, is numbered according to the codon usage table compiled by Wada *et al.* (1990); i.e. the order of an amino acid increases with the decrease of the degenerate degrees of its genetic code. If two amino acids are the same in such a degeneracy, then they are arranged in alphabetical order.

Therefore, the correlation angle $\theta(k)$ is given by

$$\theta(k)=\cos^{-1}\left\{\frac{\sum_{i=1}^{20} v_i(x)v_i(k)}{\left[\left(\sum_{i=1}^{20} v_i^2(x)\right)\left(\sum_{i=1}^{20} v_i^2(k)\right)\right]^{1/2}}\right\}. \tag{3}$$

The protein is predicted to be the j type, if $\theta(j)$ is the smallest one of the five correlation angles $\theta(\alpha)$, $\theta(\beta)$, $\theta(\alpha+\beta)$, $\theta(\alpha/\beta)$ and $\theta(\zeta)$. For convenience, hereafter, we use the indices 1, 2, 3, 4 and 5 to represent the α, β, $\alpha+\beta$, α/β and ζ type proteins, respectively. Therefore,

$$\theta(j)=\min\{\theta(1), \theta(2), \theta(3), \theta(4), \theta(5)\} \tag{4}$$

where the index j gives the predicted type of a protein.

METHOD B (CHOU, 1980)

Suppose that $\boldsymbol{v}(x)=(v_1(x), v_2(x), \ldots, v_{20}(x))$ is a point in the composition space, where $v_i(x)$ represents the frequency of amino acid i in the protein x to be predicted. The Minkowski's distance $d^M(k)$ between the point $\boldsymbol{v}(x)$ and each of the end points of the five standard vectors $\boldsymbol{v}(1)$, $\boldsymbol{v}(2)$, $\boldsymbol{v}(3)$, $\boldsymbol{v}(4)$, and $\boldsymbol{v}(5)$ is calculated, respectively, as follows

$$d^M(k)=\left\{\sum_{i=1}^{20}|v_i(x)-v_i(k)|^p\right\}^{1/p} \qquad (k=1, 2, 3, 4, 5) \tag{5}$$

where $p=1$ was used by Chou (1980). The protein is predicted to be the j type, if $d^M(j)$ is the smallest one of the five distances $d^M(1)$, $d^M(2)$, $d^M(3)$, $d^M(4)$, and $d^M(5)$.

METHOD C (NAKASHIMA *ET AL.*, 1986)

The notations used in this method are the same as those in method B. Instead of the Minkowski distance, the Euclid distance $d^E(k)$ between the point $\boldsymbol{v}(x)$ and each of the end points of the five standard vectors $\boldsymbol{v}(1)$, $\boldsymbol{v}(2)$, $\boldsymbol{v}(3)$, $\boldsymbol{v}(4)$, and $\boldsymbol{v}(5)$ is calculated as follows

$$d^E(k)=\left\{\sum_{i=1}^{20}[v_i(x)-v_i(k)]^2\right\}^{1/2} \qquad (k=1, 2, 3, 4, 5). \tag{6}$$

Similarly, the protein x is predicted to be the j type, if $d^E(j)$ is the smallest one of the five distances.

For a protein, if it is predicted as the j type by more than two of the above three methods, then it is assigned to be the j type. Otherwise, it is regarded as an uncertain structure. This is the so-called "joint prediction".

Results and Discussion

Of the 1490 human proteins, 128 proteins were predicted to be α proteins, 235 were β proteins, 169 were $\alpha+\beta$ proteins, 933 were α/β proteins, 19 were irregular proteins, and six proteins were uncertain about their folding type because none of them were predicted to be the same folding type with more than one method. The percentages of the five folding types for the 1490 human proteins are listed in Table 3, from which we can see that the α/β proteins are much more frequent than the other four types. This is a very interesting finding and its biological implication will be discussed later. Here, however, let us first give an estimation about the accuracy of the joint prediction method. This can be realized by using a set of structure-known proteins as a benchmark to inspect the predicted results one by one. The rates of correct prediction thus obtained are listed in Table 4, in which the accuracy rates for methods A and B are derived by inspecting 64 structure-known proteins, and those

TABLE 3

The percentages of the five folding types for the 1490 *human proteins*

Folding type	Number of proteins	Percentage
α	128	8·59%
β	235	15·77%
$\alpha+\beta$	169	11·34%
α/β	933	62·62%
ζ	19	1·28%
Uncertain structure†	6	0·40%

† The uncertain structure is the one whose folding type is completely inconsistent when predicted with the three different methods.

TABLE 4

The predicted results for structurally known proteins by three methods

	Rate of correct prediction					
Method	*α type*	*β type*	*$\alpha+\beta$ type*	*α/β type*	*ζ type*	*General accuracy*
A†	97·4%	80·0%	71·4%	81·3%	—	83·6%
B‡	84·2%	80·0%	78·6%	75·0%	—	79·7%
C§	87·1%	64·7%	37·0%	84·6%	50·0%	70·2%

† The prediction is performed for four types of structurally known proteins, in which there are 19 α, 15 β, 14 $\alpha+\beta$, and 16 α/β proteins (Chou & Zhang, 1992*a*).

‡ The prediction is performed for the same set of 64 proteins as in method A (Chou, 1989).

§ The prediction is performed for five types of structurally known proteins, in which there are 31 α, 34 β, 39 $\alpha+\beta$, 27 α/β, and four ζ proteins (Nakashima *et al.*, 1986). Note that the number of the irregular type to be predicted is only four, its predicted accuracy (50%) is statistically insignificant and is listed here for reference only.

for method C derived by inspecting 135 structure-known proteins. As we can see from the data of Table 4, the accuracy of our joint prediction method would be at least 80% even for a conservative estimation.

It is interesting to see from Table 3 that the percentage of α/β proteins in the 1490 human proteins is very high (62·62%). Even though the accuracy of the joint predicting method is only assumed to be 80%, the percentage of the α/β proteins would be greater than 50% of the 1490 human proteins. Of course, this does not mean that the α/β proteins would dominate in the entire set of human proteins, whose number runs around the 10^5 order of magnitude and whose codons remain to be determined. However, from a statistical point of view, the number 1490 is a "large" number. This implies that the results based on such a large number are statistically significant. Therefore, it is reasonable to envision that the α/β proteins in the entire set of human proteins might be dominant. We hope that this conclusion may be examined by the future development of protein science. On the other hand, it was shown that among 175 structurally known enzymes there are 17 with the structure of α/β barrel, i.e. 10% of all known enzymes are of the α/β barrel structure

TABLE 5

The mean frequencies of bases in the first, second and third codon position for the five folding types of proteins

Folding type	Codon position	a	c	g	t	$g+c$	g_1/g_2	Mean length
α	1	0·276	0·218	0·345	0·313	0·563	2·41	232·0
	2	0·385	0·227	0·143	0·245	0·370		
	3	0·168	0·294	0·348	0·190	0·642		
β	1	0·262	0·254	0·300	0·184	0·554	1·37	497·8
	2	0·272	0·259	0·219	0·249	0·478		
	3	0·164	0·362	0·275	0·200	0·637		
$\alpha+\beta$	1	0·301	0·211	0·299	0·189	0·510	1·49	455·7
	2	0·341	0·229	0·201	0·228	0·430		
	3	0·202	0·303	0·262	0·234	0·565		
α/β	1	0·267	0·244	0·329	0·160	0·573	1·74	399·0
	2	0·308	0·220	0·189	0·284	0·409		
	3	0·154	0·336	0·318	0·192	0·654		
ζ	1	0·205	0·250	0·259	0·268	0·509	0·81	161·9
	2	0·265	0·305	0·319	0·111	0·624		
	3	0·096	0·518	0·278	0·109	0·796		

The mean occurrence of bases A, C, G, and T for the five folding types of proteins are denoted by a, c, g and t, respectively. The subscripts 1, 2, and 3 are used to represent the codon positions, e.g. g_1 and g_2 represent the frequencies of base G in the first and second codon positions, respectively. Mean length represents the average number of the constituent amino acids for the proteins having the same folding type.

(Farber & Petsko, 1990), implying that such a special structure of α/β folding type has already occupied a remarkable percentage of structure-known enzymes. It is rational, therefore, to deduce that the α/β proteins play a vital role in the metabolic process of the human body.

It was shown that the amino acid composition of a protein is mainly determined by the occurrence frequencies of bases in the first and second codon position for the coding sequence (Zhang & Zhang, 1991*b*). Since each folding type of protein has its characteristic amino acid composition (Table 1), it is expected that the base compositions in the first and second codon position have an important influence on the folding types of proteins. The occurrence frequencies of bases in the first, second and third codon position for each folding type of protein has been calculated, respectively. Table 5 lists the mean frequencies of them for each set of proteins. We can see from this table that the structural classes of proteins are dependent on the mean frequencies of bases in the first and second codon position. The G+C % in the first codon position for the five folding types of proteins are roughly equal to each other. However, the G+C % in the second codon position are quite different for different folding types of proteins. For such an index, the α type protein has the smallest value (0.37), the β type protein has the largest value (0·478), the $\alpha+\beta$ and α/β proteins have values in between, and the irregular folding type of protein has an anomalously large

value (0·624). The ratio of the frequency of base G in the first codon position to that in the second codon position is a useful statistical quantity (Trifonov, 1987). As we can see, this ratio for the irregular structure of proteins has an anomalously small value (0.81). In addition to this, the ratio for the α type has the largest value (2·41), the ratio for the β type has the smallest value (1·37), and those for the $\alpha+\beta$ and α/β types have values in between. Based on this observation, we suggest that this ratio may be regarded as an index to characterize the folding types of the proteins. In other words, the folding type of a protein is strongly dependent on the ratio of the frequency of base G in the first codon position to that in the second codon position.

According to the genetic code, there is less connection between the amino acid composition of a protein and the occurrence frequencies of bases in the third codon position in this protein coding sequence. The reason is that most of the amino acids are encoded by the degenerate codons. However, we still think that it is worthwhile to study the possible and weak correlation between the folding types of 1490 human proteins and the occurrence frequencies of bases in the third codon position in the DNA sequences coding for these proteins. The distribution of base composition in the third codon position is studied by a diagrammatic technique (Zhang & Zhang, 1991*a*, *b*; Chou & Zhang, 1992*b*). In Fig. 1(a) and (b), there are 933 points in both diagrams, in which each point represents an α/β protein as predicted from the 1490 human proteins. In Fig. 2 there are a total of $128+235+169+19=551$ points in both diagrams (a) and (b), representing the 128 predicted α proteins, 235 β proteins, 169 $\alpha+\beta$ proteins and 19 irregular proteins, respectively. The points representing the different folding types of proteins are distinguished by different symbols. As we can see from these figures the distribution regions of points for the five folding types of proteins severely overlap. This implies that generally the folding type of a protein does not depend on the occurrence frequencies of bases in the third codon position. However, by a detailed observation of Fig. 2(a) and (b) we find that for most of the α proteins the representing points are distributed in the region of $y<0$. According to the diagram principle (Zhang & Zhang, 1991*a*, *b*), this means that $a+c<1/2$, where the relative occurrence frequencies of bases A, C, G and T in the third codon position are denoted by a, c, g and t, respectively. Furthermore, the points representing the irregular folding type of proteins gather together and are distributed in the region of the second quadrant in Fig. 2(a) and of the fourth quadrant of Fig. 2(b). This implies that $a+c>1/2$, $g+c>1/2$ and $t+c>1/2$. Since the points of different folding types are situated near either of the two diagonals of the square in the diagram (Chou & Zhang, 1992*b*), we have approximately $a=g=t$. These coding characteristics in the third codon position are necessary conditions for the relevant folding types of proteins, but they are not sufficient conditions. This is because the points representing different folding types of proteins overlap severely.

The degeneracy phenomenon between amino acids and their genetic codons is a very interesting phenomenon, although little is yet known about its essence. Is there anything deeper hidden behind the phenomenon? According to quantum mechanics, the degenerate energy level of an atom will be split when it is in the environment of an electrostatic or magnetic field. Are there any "similar" effects for the degenerate

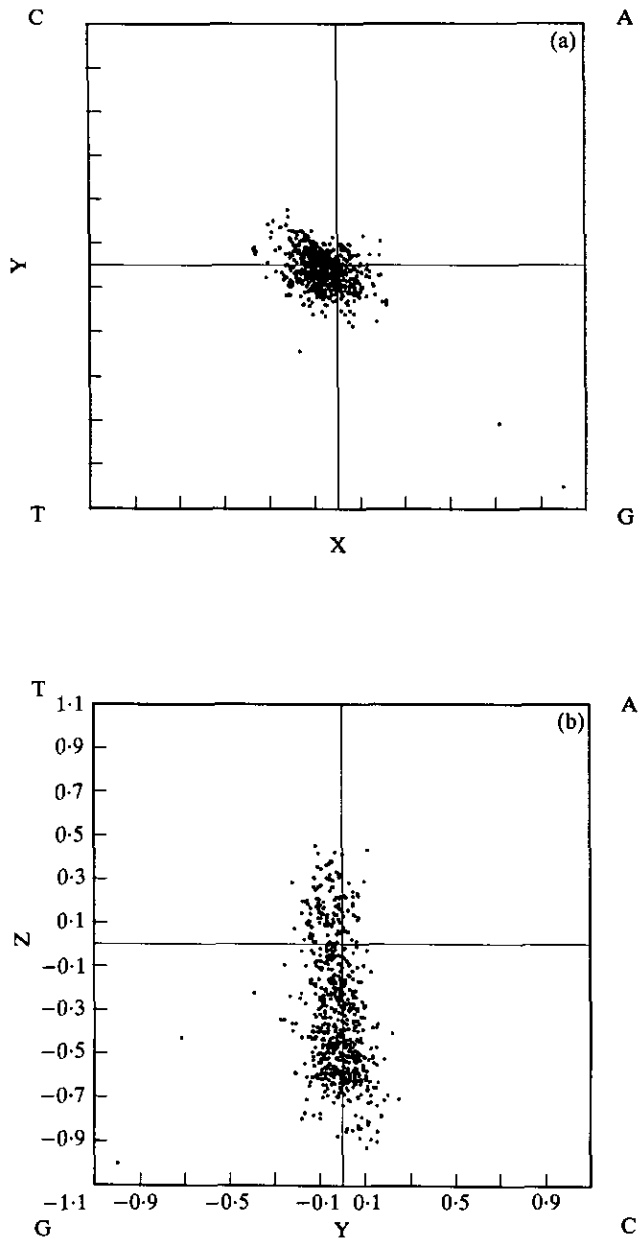

FIG. 1. The distribution of base composition in the third codon position for the 933 predicted α/β proteins. Suppose that the relative frequencies of bases A, C, G and T (or U) in the third codon position are denoted by a, c, g and t, respectively, then $x=2(a+g)-1$, $y=2(a+c)-1$, $z=2(a+t)-1$. Figure 1 (a) is the distribution diagram on the x–y plane, (b) on the y–z plane. Each point represents one protein. Therefore, there are a total of 933 points in each diagram. For details about the figures see Zhang & Zhang (1991*a*, *b*).

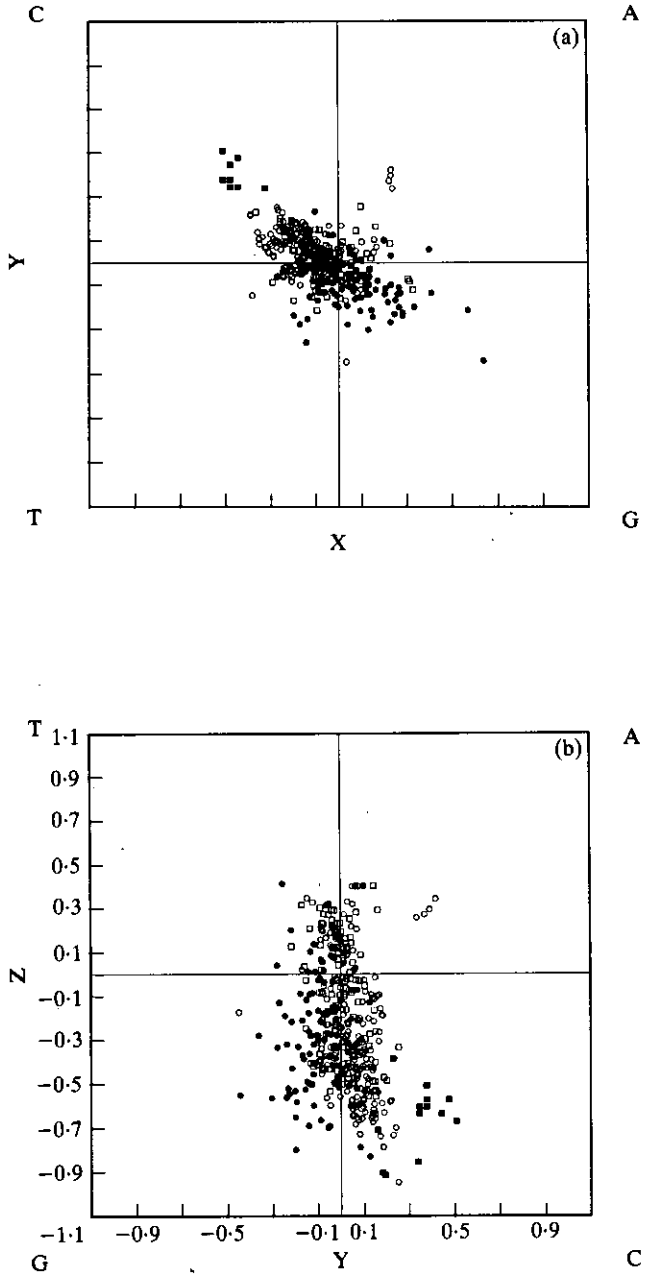

FIG. 2. The diagrammatic representation of base composition in the third codon position for the 128 predicted α, 236 β, 169 $\alpha+\beta$ and 19 irregular proteins: (a) on the x–y plane, and (b) on the y–z plane. There are in total $128+235+169+19=551$ points in each diagram. The α proteins are represented by (●), β by (○), $\alpha+\beta$ by (□) and the irregular proteins by (■). See the legend of Fig. 1 for a further explanation.

codons, such as those associated with the molecular mutation or replication of proteins? This is obviously an interesting problem, and certainly worthy of further investigation.

The average chain lengths of proteins for the five folding types are calculated and listed in Table 5. It can be seen that, next to the irregular folding type, the average length of the α type has the smallest value. This is in agreement with the observation by Chou (1989). According to his analysis of 64 structurally known proteins in which there are 19 α, 15 β, 14 $\alpha+\beta$ and 16 α/β proteins, the average chain length of the 19 α proteins has the smallest value, 129. Therefore, it is reasonable to say that of the four folding types of proteins (α, β, $\alpha+\beta$, and α/β) the α type has the shortest average chain length.

In conclusion, our study is based on the following principle. The folding type or the structural class of a protein is related to its amino acid composition (Chou, 1980, 1989; Nakashima *et al.*, 1986; Chou & Zhang, 1992*a*). The amino acid composition of a protein is determined by its codon usage (Wada *et al.*, 1990). Therefore, the folding type or the structural class of a protein is related to its codon usage in the DNA sequence coding for this protein. In this paper the detailed codon characteristics of five folding types of proteins have been studied for 1490 human proteins. The overall results are listed in Table 5. These results provide an intuitive picture about the relationship between protein folding type and the codon usage. Although the method that we use to predict the folding type of a protein from merely its amino acid composition is not an absolutely accurate one, nonetheless with an accuracy of at least 80%, the results thus obtained are still worth noting. So far as we know, it is the first time that the folding types of so many proteins have been studied based on their codon usage in the coding sequence.

Since the three-dimensional structures of most of the 1490 human proteins have not yet been determined, our predicted results about their folding types could play a significant role in stimulating the study of this area from both an experimental and theoretical point of view.

REFERENCES

AOTA, S., GOJOBORI, T., ISHIBASHI, F., MARUYAMA, T. & IKEMURA, T. (1988). *Nucl. Acids Res.* **16**, r315–r402.

CHOU, K. C. & ZHANG, C.-T. (1992*a*). *Eur. J. Biochem.* **207**, 429–433.

CHOU, K. C. & ZHANG, C. T. (1992*b*). AIDS *Res. Hum. Retroviruses* **8**, 1967–1976.

CHOU, P. Y. (1980). In: *Abstract of Papers, Part I, Second Chemical Congress of the North American Continent*, Las Vegas.

CHOU, P. Y. (1989). In: *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.) pp. 549–586. New York: Plenum Press.

FARBER, G. K. & PETSKO, G. A. (1990). *Trends Biochem. Sci.* **15**, 228–234.

GRANTHAM, R., GAUTIER, C., GOUY, M., MERCIER, R. & PAVE, A. (1980). *Nucl. Acids Res.* **8**, r48–r62.

GRANTHAM, R., GAUTIER, C., GOUY, M., JACOBZONE, M. & MERCIER, R. (1981). *Nucl. Acids Res.* **9**, r43–r74.

LEVITT, M., CHOTHIA, T., AOTA, S. & IKEMURA, T. (1976). *Nature, Lond.* **261**, 552–558.

MARSHALL, R. E., CASKEY, C. T. & NIRENBERG, M. (1967). *Science* **155**, 820–825.

MARUYAMA, T., GOJOBORI, T., AOTA, S. & IKEMURA, T. (1986). *Nucl. Acids Res.* **14**, r151–r197.

NAKASHIMA, H., NISHIKAWA, K. & OOI, T. (1986). *J. Biochem.* **99**, 153–162.

NIRENBERG, M. W. (1963). *Sci. Am.* **208**, 80–94.

RICHARDSON, J. S. & RICHARDSON, D. C. (1989). In: *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.) pp. 1–98. New York: Plenum Press.
TRIFONOV, E. N. (1987). *J. molec. Biol.* **194,** 643–652.
WADA, K., AOTA, S., TSUCHIYA, R., ISHIBASHI, F., GOJOBORI, T. & IKEMURA, T. (1990). *Nucl. Acids Res.* **18,** r2367–r2411.
ZHANG, C.-T. & ZHANG, R. (1991*a*). *Int. J. Biol. Macromol.* **13,** 45–49.
ZHANG, C.-T. & ZHANG, R. (1991*b*). *Nucl. Acids Res.* **19,** 6313–6317.