Graphic Analysis of Codon Usage Strategy in 1490 Human Proteins

Chun-Ting Zhang[1,2] and Kuo-Chen Chou[1,3]

Received January 15, 1993

The frequencies of bases A (adenine), C (cytosine), G (guanine), and T (thymine) occurring in codon position i, denoted by a_i, c_i, g_i, and t_i, respectively ($i=1, 2, 3$), have been calculated and diagrammatized for the 1490 human proteins in the codon usage table for primate genes compiled recently. Based on the characteristic graphs thus obtained, an overall picture of codon base distribution has been provided, and the relevant biological implication discussed. For the first codon position, it is shown in most cases that G is the most dominant base, and that the relationship $g_1 > a_1 > c_1 > t_1$ generally holds true. For the second codon position, A is generally the most dominant base and G is the one with the least occurrence frequently, with the relationship of $a_2 > t_2 > c_2 > g_2$. As to the third codon position, the values of $g_3 + c_3$ vary from 0.27 to 1, roughly keeping the relationship of $c_3 > g_3 > a_3 = t_3$ for the majority of cases. Interestingly, if the average frequencies for bases A, C, G, and T are defined as $\bar{a} = (a_1 + a_2 + a_3)/3$, $\bar{c} = (c_1 + c_2 + c_3)/3$, $\bar{g} = (g_1 + g_2 + g_3)/3$, and $\bar{t} = (t_1 + t_2 + t_3)/3$, respectively, we find that $\bar{a}^2 + \bar{c}^2 + \bar{g}^2 + \bar{t}^2 < \frac{1}{3}$ is valid almost without exception. Such a characteristic inequality might reflect some inherent rule of codon usage, although its biological implications is unclear. An important advantage by introducing graphic methods is to make it possible to catch essential features from a huge amount of data by a direct and intuitive examination. The method used here allows one to see means and variances, and also spot outliers. This is particularly useful for finding and classifying similarity patterns and relationships in data sets of long sequences, such as DNA coding sequences. The current method also holds a great potential for the study of molecular evolution from the viewpoint of genetic code whose data have been accumulated rapidly and are to continue growth at a much faster pace.

KEY WORDS: Codon position; DNA bases; characteristic inequality; mapping point.

1. INTRODUCTION

Grantham and his colleagues (1980, 1981) reported and analyzed the codon usages in a total of 161 genes then available. These studies have been recognized as pioneering work in this area. Since then, the number of known DNA coding sequences has grown quickly. The codon usage of 1638, 3681, and 11,415 genes were compiled and analyzed by Ikemura and his colleagues in 1986, 1988, and 1990, respectively (Maruyama *et al.*, 1986; Aota *et al.*, 1988; Wata *et al.*, 1990). It can be envisioned that the size of the database will grow at an even faster pace in the near future. Facing this situation, we find that the analysis of these data obviously falls far behind its development. Therefore, it is necessary to speed up data analysis. With the rapid accumulation of DNA sequence data, including the advent of the human genome initiative, there is an increasing need for expressing and classifying similarity patterns and relationships in an efficient and intuitive way for data sets of long sequences. In view of this, in this paper we shall introduce a simple graphic technique, and use it to analyse the codon usage strategy for the 1490 human proteins in the

[1] Computational Chemistry, Upjohn Research Laboratories, Kalamazoo, Michigan 49001.

[2] On sabbatical leave from Department of Physics, Tianjin University, Tianjin, China.

[3] To whom all correspondence should be addressed.

0277-8033/93/0600-0329$07.00/0 © 1993 Plenum Publishing Corporation

codon usage table compiled by Wata *et al.* (1990) for primate genes. As is well known, once the codon usage is known for a protein, the frequencies of occurrence of four DNA bases in the first, second, and third codon positions may be calculated. However, we have $4 \times 3 = 12$ such data for each protein. For the 1490 human proteins, the number of total data will be $12 \times 1490 = 17{,}880$! It would take several pages to print out these data! It is almost impossible to analyze these data directly. Under such circumstances, using the graphic technique described below is particularly effective. Its main merit is reflected by the fact that a lot of data can be summarized in a readily perceivable form by few diagrams. By looking at them, one may easily and quickly catch the essential features, drawing some overall conclusions about the distribution of bases in the three codon positions. The present study is dedicated to reveal the codon usage strategy in the 1490 human proteins in terms of the graphic method.

2. METHODS

Let the frequency of occurrence of base A (adenine) in the ith ($i = 1, 2, 3$) codon position be denoted by a_i. Similar symbols are used for bases C (cytosine), G (guanine), and T (thymine). Obviously, we have

$$\left.\begin{array}{l} a_i + c_i + g_i + t_i = 1 \\ 0 \le a_i,\ c_i,\ g_i,\ t_i \le 1 \end{array}\right. \quad (i = 1, 2, 3) \qquad (1)$$

Since the following formulation is generally valid regardless of which one of the three codon positions is refereed, for brevity the subscript i will be omitted below unless those cases in which a special reference mark is needed for distinction. It is implied from Eq. (1) that, of the above four real numbers a, c, g, and t, only three are independent. Therefore, they can be expressed in terms of only three independent variables, x, y, and z, as formulated by the following equation:

$$\begin{bmatrix} a \\ c \\ g \\ t \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} \qquad (2)$$

The above matrix equation can also be written as

$$x = 2(a+g) - 1 = 1 - 2(t+c) \qquad (3a)$$

$$y = 2(a+c) - 1 = 1 - 2(t+g) \qquad (3b)$$

$$z = 2(a+t) - 1 = 1 - 2(g+c) \qquad (3c)$$

where x, y, and z are the coordinates in a Cartesian coordinate system. Therefore, the occurrence frequencies for A, C, G, and T can be uniquely defined in a three-dimensional codon space by one point, the so-called "codon mapping point."

Because the distribution of points in a three-dimensional space can be completely described by two orthogonal coordinate planes, we shall use X–Y and Y–Z planes hereafter. In the X–Y plane (Fig. 1a), according to Eq. (3) the allowed regions for x, y, and

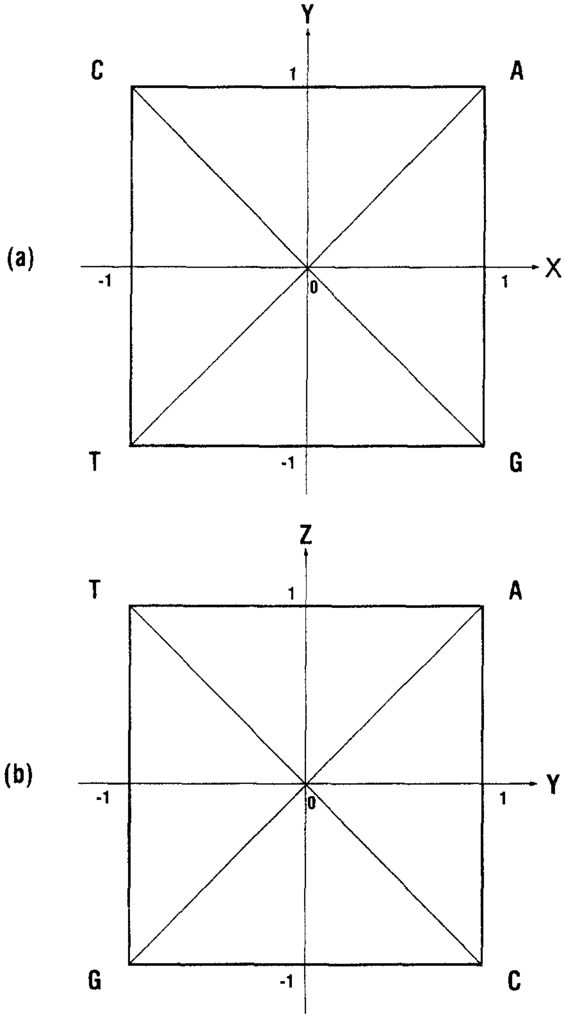

Fig. 1. Projection of the mapping point space according to Eq. (2) to (a) the X–Y coordinate plane, and (b) the Y–Z coordinate plane. The projected areas in these two planes are a regular square whose side length is 2. The point A is the vertex of the square in the X–Y plane, corresponding to the coding sequence along which a given codon position, say the ith ($i = 1$, 2, or 3), is purely occupied by base A (i.e., $a_i = 1$). The point C is another such vertex, and so forth. For simplicity, however, all the subscripts referring to the condon positions and coordinate planes are omitted but they can be easily identified by the context.

z are from -1 to $+1$. Therefore, the distribution of the projected points in each coordinate plane is contained with a square with sides of 2 (Fig. 1a). If the base in the ith ($i=1$, 2, or 3) codon position for a protein is always base A, i.e. [$a=1$, $c=g=t=0$, according to Eqs. (3a) and (3b)], then the projection of the corresponding mapping point to the X–Y plane will be the vertex A with $x=1$ and $y=1$ (Fig. 1a). Likewise, the vertices C, G, and T have the similar implication, and they, together with A, actually form the four vertices of the square in Fig. 1a. Furthermore, the shorter the distance of a point from the vertex, say A, the greater the component of base A in the corresponding codon position. Therefore, using the symbol $\Leftrightarrow$ to express the mapping relationship, we have

In X–Y plane:

$$\mathrm{A} \Leftrightarrow (a=1, c=g=t=0);$$

$$\mathrm{C} \Leftrightarrow (c=1, a=g=t=0);$$

$$\mathrm{G} \Leftrightarrow (g=1, a=c=t=0);$$

$$\mathrm{T} \Leftrightarrow (t=1, a=c=g=0);$$

$$\text{Line AG} \Leftrightarrow (a+g=1, c=t=0);$$

$$\text{Line GT} \Leftrightarrow (g+t=1, a=c=0);$$

$$\text{Line TC} \Leftrightarrow (c+t=1, g=a=0);$$

$$\text{Line CA} \Leftrightarrow (a+c=1, g=t=0);$$

$$x=0 \Leftrightarrow (a+g=\tfrac{1}{2});$$

$$x>0 \Leftrightarrow (a+g>\tfrac{1}{2});$$

$$x<0 \Leftrightarrow (c+t>\tfrac{1}{2});$$

$$y=0 \Leftrightarrow (a+c=\tfrac{1}{2});$$

$$y>0 \Leftrightarrow (a+c>\tfrac{1}{2});$$

$$y<0 \Leftrightarrow (g+t>\tfrac{1}{2});$$

$$\text{Line AOT} \Leftrightarrow (g=c);$$

$$\triangle\text{AGT} \Leftrightarrow (g>c);$$

$$\triangle\text{ATC} \Leftrightarrow (c>g);$$

$$\text{Line GOC} \Leftrightarrow (a=t);$$

$$\triangle\text{AGC} \Leftrightarrow (a>t);$$

$$\triangle\text{GTC} \Leftrightarrow (t>a);$$

$$\triangle\text{AOG, called region I} \Leftrightarrow (a>t \text{ and } g>c);$$

$$\triangle\text{AOC, called region II} \Leftrightarrow (a>t \text{ and } c>g);$$

$$\triangle\text{COT, called region III} \Leftrightarrow (t>a \text{ and } c>g);$$

$$\triangle\text{GOT, called region IV} \Leftrightarrow (t>a \text{ and } g>c);$$

$$\text{Point } 0 \Leftrightarrow (a=c=g=t=\tfrac{1}{4})$$

Now let us turn to the Y–Z projecting plane. Referring to Fig. 1b and according to Eqs. (3b) and (3c), we have

In Y–Z plane:

$$\mathrm{A} \Leftrightarrow (a=1, c=g=t=0);$$

$$\mathrm{C} \Leftrightarrow (c=1, a=g=t=0);$$

$$\mathrm{G} \Leftrightarrow (g=1, a=c=t=0);$$

$$\mathrm{T} \Leftrightarrow (t=1, a=c=g=0);$$

$$\text{Line AC} \Leftrightarrow (a+c=1, g=t=0);$$

$$\text{Line CG} \Leftrightarrow (c+g=1, a=t=0);$$

$$\text{Line GT} \Leftrightarrow (g+t=1, a=c=0);$$

$$\text{Line TA} \Leftrightarrow (a+t=1, c=g=0);$$

$$y=0 \Leftrightarrow (a+c=\tfrac{1}{2});$$

$$y>0 \Leftrightarrow (a+c>\tfrac{1}{2});$$

$$y<0 \Leftrightarrow (g+t>\tfrac{1}{2});$$

$$z=0 \Leftrightarrow (g+c=\tfrac{1}{2});$$

$$z>0 \Leftrightarrow (g+c<\tfrac{1}{2});$$

$$z<0 \Leftrightarrow (g+c>\tfrac{1}{2});$$

$$\text{Line AOG} \Leftrightarrow (c=t);$$

$$\triangle\text{ACG} \Leftrightarrow (c>t);$$

$$\triangle\text{AGT} \Leftrightarrow (t>c);$$

$$\text{Line COT} \Leftrightarrow (a=g);$$

$$\triangle\text{ACT} \Leftrightarrow (a>g);$$

$$\triangle\text{CGT} \Leftrightarrow (g>a);$$

$$\triangle\text{AOC, called region I} \Leftrightarrow (a>g \text{ and } c>t);$$

$$\triangle\text{AOT, called region II} \Leftrightarrow (a>g \text{ and } t>c);$$

$$\triangle\text{TOG, called region III} \Leftrightarrow (g>a \text{ and } t>c);$$

$$\triangle\text{COG, called region IV} \Leftrightarrow (g>a \text{ and } c>t);$$

$$\text{Point } 0 \Leftrightarrow (a=c=g=t=\tfrac{1}{4})$$

The conclusions derived from the distribution of points in the X–Y plane will never contradict those from the corresponding distribution in the Y–Z plane. On the contrary, the observations from the two projecting planes always complement each other. Because the mapping points are distributed in a three-dimensional space, the complete results about the distribution can be obtained by observing the distribution of projecting points in both X–Y and Y–Z coordinate planes.

The whole process of graphic representation can be summarized as follows. (a) Calculate the frequencies of bases in the first, second, and third codon positions, respectively, for the DNA sequences coding for the 1490 human proteins according to the codon usage table (Wata *et al.*, 1990). (b) Map the data thus obtained onto a three-dimensional space according to Eq. (3); for each of the codon positions there are 1490 mapping points, which are further displayed on the X–Y coordinate planes for the convenience of intuitive analysis. The program for this type of graphic analysis is available upon request.

Some conclusions may be drawn by studying the distribution of the points on these coordinate planes, as illustrated below.

3. RESULTS AND DISCUSSION

3.1. The Distribution of Bases in the First Codon Position

Refer to Fig. 2a, where the mapping points are projected onto the X_1–Y_1 plane. Note that almost all points are situated in the region of $x_1>0$. According to the description in the last section, this means that $a_1+g_1>\frac{1}{2}$, or the purine bases A and G are dominant in the first codon position for the 1490 human proteins. Furthermore, by drawing two diagonals (not shown in this figure), we find the points are almost situated in region I (i.e., $a_1>t_1$ and $g_1>c_1$). The 1490 points are roughly symmetrical with respect to the Y_1 axis. Now turn to Fig. 2b. Note that most of the points are in the region of $z_1<0$ (i.e., $g_1+c_1>\frac{1}{2}$), and the points are mainly distributed in region IV (i.e., $g_1>a_1$ and $c_1>t_1$). Also, the points are roughly symmetrical with respect to the Y_1 axis. Summarizing the results observed in Fig. 2a and b, we conclude that the purine bases A and G are dominant, and that G is the most dominant base in the first codon position in most cases. Generally speaking, we have $g_1>a_1>c_1>t_1$, and $a_1+c_1\approx\frac{1}{2}$, $g_1+t_1\approx\frac{1}{2}$.

3.2. The Distribution of Bases in the Second Codon Position

Refer to Fig. 3a first. Most of the points are situated in the region of $Y_2>0$ (i.e., $a_2+c_2>\frac{1}{2}$), and at the same time the points are in region II (i.e., $a_2>t_2$ and $c_2>g_2$). Now turn to Fig. 3b. The points are nearly in the region of $\triangle$ACT (i.e., $a_2>g_2$). Most of the points are in the region of $Z_2>0$ (i.e., $g_2+c_2<\frac{1}{2}$), and nearly all the points are in region II (i.e., $t_2>c_2$, $a_2>g_2$). Summarizing the results observed in Fig. 3a and b, we conclude that A is the most dominant base and G is the least dominant base in the second codon position for most of the 1490 human proteins. Generally, we have $a_2>t_2>c_2>g_2$, and $a_2+c_2>\frac{1}{2}$, $g_2+c_2<\frac{1}{2}$.

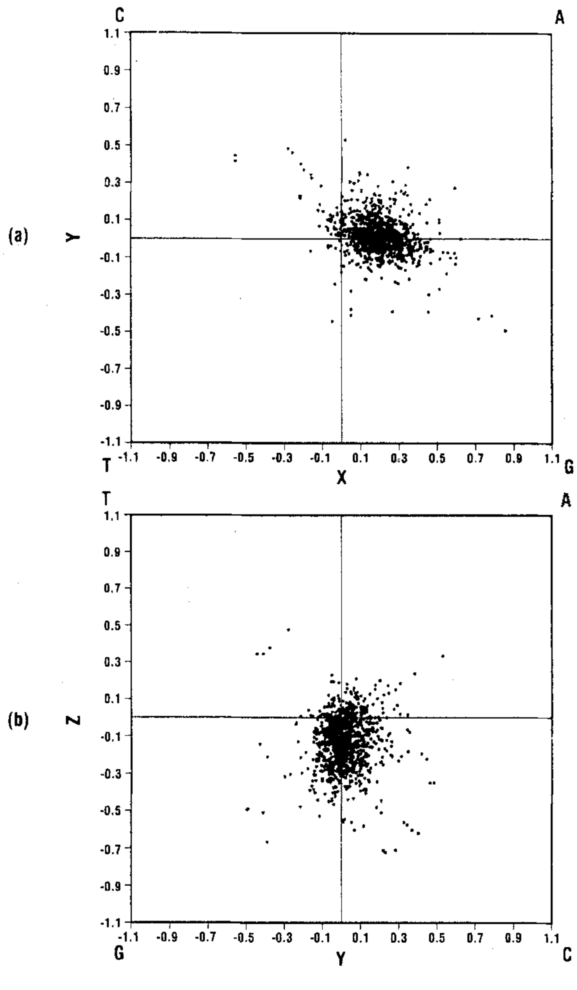

Fig. 2. The occurrence frequencies of the four bases at the first codon position for the 1490 human protein coding sequences are mapped according to Eq. (3) to (a) the X–Y coordinate plane, and (b) the Y–Z coordinate plane. Each of such sequences corresponds to a mapping point, and there are totally 1490 points on each of the two planes. See legend to Fig. 1 for further explanation.

3.3. The Distribution of Bases in the Third Codon Position

Refer to Fig. 4a. Note that most of the points are in the region of $X_3<0$ (i.e., $c_3+t_3>\frac{1}{2}$)—namely, the pyrimidine bases are dominant in the third position for the coding sequences for most of the 1490 human proteins. Also note that there is a tendency that the points are distributed along the diagonal

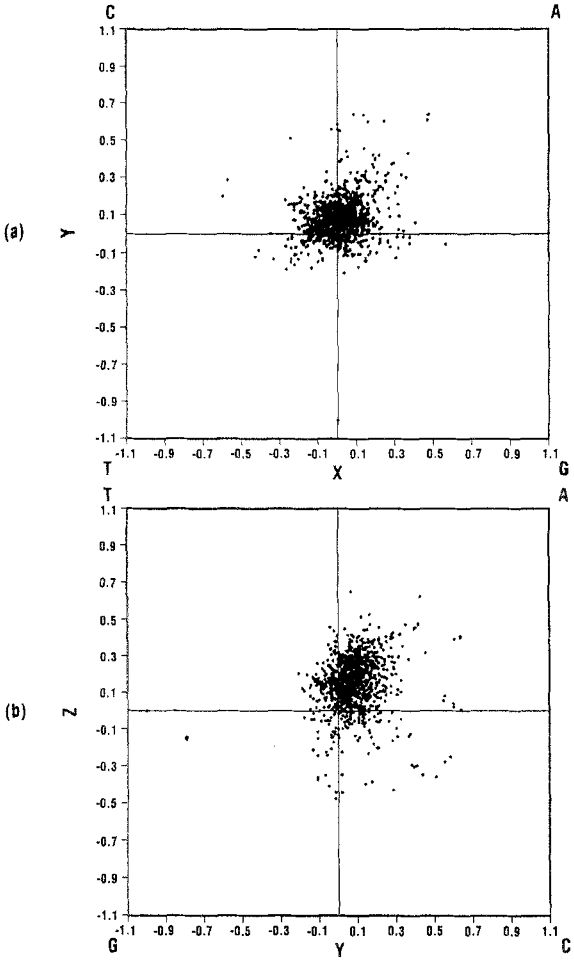

Fig. 3. The occurrence frequencies of the four bases at the second codon position for the 1490 human protein coding sequences are mapped according to Eq. (3) to (a) the X-Y coordinate plane, and (b) the Y-Z coordinate plane. See legend to Fig. 1 for further explanation.

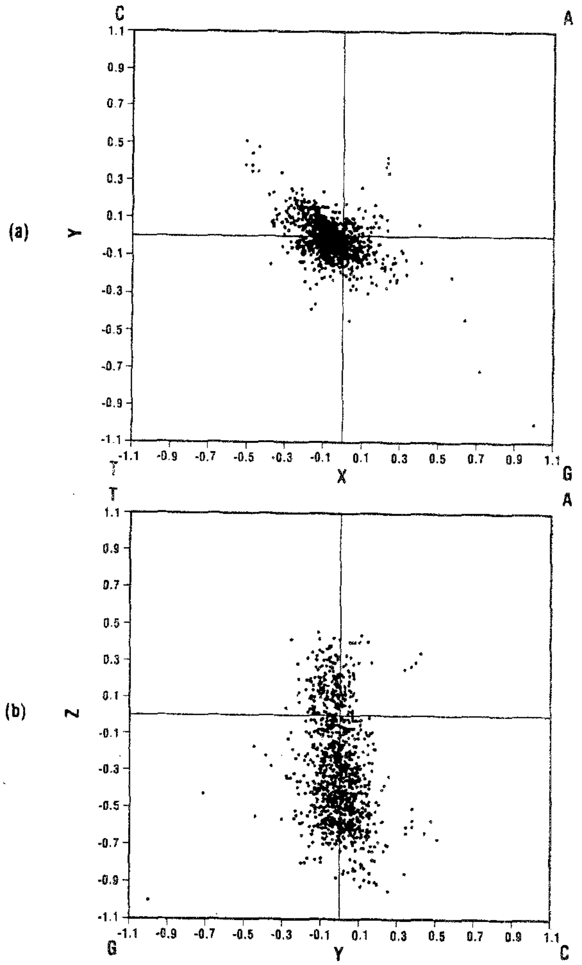

Fig. 4. The occurrence frequencies of the four bases at the third codon position for the 1490 human protein coding sequences are mapped according to Eq. (3) to (a) the X-Y coordinate plane, and (b) the Y-Z coordinate plane. See legend to Fig. 1 for further explanation.

GOC; thus, we have $a_3 = t_3$ approximately. Now turn to Fig. 4b. Note that the distribution is scattered in a much larger region than those seen before. The points are distributed in an elliptic column-like region in a three-dimensional space. This implies that a large variety of choices are available for the bases in the third codon position for the DNA sequences coding for different human proteins. Most of the points are situated in the region of $Z_3 < 0$ (i.e., $g_3 + c_3 > \frac{1}{2}$). The largest value of $g_3 + c_3$ is 100%, as shown by the fact that there is a point which is located just at the vertex G in Fig. 4b. The largest value of Z_3 is 0.4537, which corresponds to the smallest value of $g_3 + c_3$ being equal to 0.27. Note that most of the points are in region IV (i.e., $c_3 > t_3$ and $g_3 > a_3$). Summarizing the results seen in Fig. 4a and b, we find that $g_3 + c_3$ varies from 0.27 to 1, keeping approximately $c_3 > t_3$, $g_3 > a_3$ and $a_3 = t_3$. Generally speaking, we have $c_3 > g_3 > a_3 = t_3$ in most cases.

More than 10 years ago, Grantham (1980) pointed out that mRNA sequences contain other information than that necessary for encoding proteins. The other information is mainly in the degenerate bases of the third position. It is well known that the choice of bases in the third position is species-specific. Such a fact is called a "genome hypothesis" (Grantham, 1980) or a "codon dialect" (Ikemura, 1985). So, the comparison of Fig. 4a and b with those for other species is meaningful. According to the genome hypothesis or the theory of codon dialect, it is reasonable to expect that the distribution patterns of points in Fig. 4a (X_3–Y_3 projection plane) and Fig.

4b (Y_3–Z_3 projection plane) for taxonomically related species should be similar, but be different for distant species. Consequently, graphs such as Fig. 4 hold a great potential for the study of evolution and classification of organisms.

On the other hand, it was shown that the bias in codon choice within genes in a single species appears related to the level of expression of the protein encoded by that gene (Murray *et al.*, 1989). The bias is mainly in the use of particular bases in the third position. It was shown that the highly expressed plant genes share a similar extreme preference for G+C in the third position of codons (Murray *et al.*, 1989). In graphic terminology, those points with high values of g_3+c_3 should be situated in the lower part of Fig. 4b. It is expected that the genes associated with the points in the extreme low part of Fig. 4b might be of the highly expressed proteins.

Recently, Ikemura and Wata (1991) have studied the chromosome locations of the evidently (g_3+c_3)-rich, as well as the (a_3+t_3)-rich human genes. They have found that a major portion of the (g_3+c_3)-rich human genes was found to be on special subsets of R-bands (T-bands and/or terminal R-bands). Those genes with large values of a_3+t_3, however, were mainly on G-bands or non-T-type internal R-bands (Ikemura and Wata, 1991). According to graphic terminology, the points with large values of g_3+c_3 are situated in the lower part of the set of points in Fig. 4b; while those with large values of a_3+t_3 are in the upper part of the set of points. Therefore, the distribution of points in Fig. 4b may reflect to some extent the locations of genes in the chromosomes.

3.4. The Distribution of Bases in the Overall Coding Sequences

The overall average of frequencies of bases A, C, G, and T over the coding sequences for the 1490 human proteins have been calculated. The standard deviations have also been calculated. These results are listed in Table I, from which we can see that the overall average occurrence frequency for G+C content is 0.54, while the overall G+C in the human genome is roughly 0.40 (Ikemura and Wata, 1991). This means that there are some other sequences with rich A and T bases in the human genomic DNA.

Taking the average of frequencies for bases A, C, G, and T over the three codon positions, we obtain the corresponding average frequencies $\bar{a}$, $\bar{c}$, $\bar{g}$, and $\bar{t}$. The average distribution of bases is shown in Fig. 5a and b, respectively. The overall appearance of the two

Table I. Average Frequencies of Bases A, C, G, and T over 1490 Human Protein Coding Sequences and Their Standard Deviation

Base	A	C	G	T
Average frequency	0.2475	0.2678	0.2743	0.2099
SD	0.0479	0.0520	0.0412	0.0400

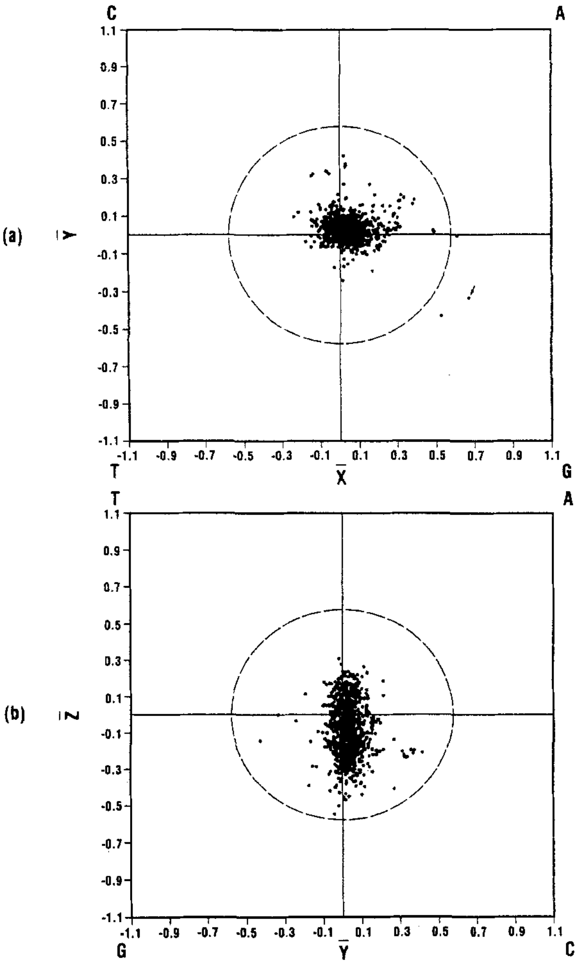

Fig. 5. The overall frequencies of bases for each of the 1490 sequences are defined as follows: $\bar{a}=(a_1+a_2+a_3)/3$, $\bar{c}=(c_1+c_2+c_3)/3$, $\bar{g}=(g_1+g_2+g_3)/3$, $\bar{t}=(t_1+t_2+t_3)/3$. These frequencies are mapped and projected to (a) the $\bar{X}$–$\bar{Y}$ plane, and (b) the $\bar{Y}$–$\bar{Z}$ plane. The circle in each of the two diagrams has a radius of $1/\sqrt{3}$ = 0.577. If the projection points of a protein coding sequence to both the $\bar{X}$–$\bar{Y}$ and the $\bar{Y}$–$\bar{Z}$ planes are inside the circle, then the characteristic inequality of Eq. (4) is valid for this coding sequence; otherwise, it is invalid. The point marked by an arrow in panel (a) represents two overlapped points corresponding to two different 3-residue peptides with, however, the same average base occurrence frequencies.

graphs is that all the points are nearer to the origin point 0 than the corresponding ones seen in the previous graphs. It is intriguing to note that the following relation

$$\bar{a}^2+\bar{c}^2+\bar{g}^2+\bar{t}^2<\tfrac{1}{3} \tag{4}$$

holds true for 1486 of the 1490 human protein coding sequences. Equation (4) can be further illustrated through Fig. 5a and b, where, for clarity, a circle is drawn that is centered at 0 and with radius $1/\sqrt{3}=0.577$. If a mapping point in both Fig. 5a and b is within the circle, then Eq. (4) is valid for the corresponding sequence; otherwise, Eq. (4) is invalid. As we can see, there are three points located outside of the circle in Fig. 5a. Of these three points, the one marked with an arrow represents two overlapped points, which are corresponding to two different coding sequences with, however, the same average base occurrence frequencies. Therefore, in Fig. 5a there are actually four points located outside of the circle. However, careful examination would reveal that, of the four points, only one is corresponding to the coding sequence of a protein with 102 amino acid residues, but the other three (including the two overlapped points) are corresponding to those of very short peptides with only 3, 3, and 8 amino acid residues, respectively. They cannot be counted as proteins. Therefore, there is only one protein whose mapping point exceeds beyond the circle limit, violating the inequality of Eq. (4). The biological meaning implied in such a characteristic inequality is still not clear, although it holds true almost unexceptionally. What inherent relationship does it reflect? Is there any important biological implication hidden in the mathematical inequality? Further investigation into these problems from the viewpoint of codon usage strategy would certainly be rewarding.

4. CONCLUSIONS

The frequencies of bases A, C, G, and T in the first, second, and third codon positions for the DNA sequences coding for 1490 human proteins have been calculated. These data are then mapped onto the points in a three-dimensional space, followed by being displayed on the X–Y and Y–Z coordinate planes for an intuitive analysis and study. It is shown that at the first codon position for the majority of cases, G is the most dominant base with the general relationship $g_1>a_1>c_1>t_1$. At the second codon position, for the majority of cases A is the most dominant base and G is the one with the lowest frequency. Generally speaking, for the second codon position of the 1490 human proteins, we find $a_2>t_2>c_2>g_2$. As to the third codon position, the values of g_3+c_3 vary from 0.27 to 1, and in most cases $c_3>g_3>a_3=t_3$. The average frequencies, $\bar{a}$, $\bar{c}$, $\bar{g}$, and $\bar{t}$, of the four bases at the three codon positions over the 1490 human proteins have also been calculated. A limitation has been observed that the inequality $\bar{a}^2+\bar{c}^2+\bar{g}^2+\bar{t}^2<\frac{1}{3}$ holds true almost unexceptionally for the 1490 human protein coding sequences, and hence it can be termed as the characteristic inequality of codon usage. This is an interesting finding and might reflect some important law of nature, although at present its biological meaning is not quite clear yet.

An obvious benefit of introducing the graphic representation is to make it possible to catch the essential features from a huge amount of data by direct and intuitive examination. This is particularly useful in performing cluster analysis, as demonstrated in this paper. In view of the fact that the data of genetic code have been accumulated rapidly and the pace of such growth is to continue to increase, it can be anticipated that the graphic method as used here in analyzing the codon patterns of the 1490 human proteins and revealing their biological implications will hold a great potential for the study of molecular evolution in genetic terms.

ACKNOWLEDGMENTS

We are indebted to Prof. T. Ikemura for kindly sending us the tape of the 1490 human protein codon usage data. We are also very grateful to Dr. Wei-Zhu Zhong for her help in preparing Figs. 2–5.

REFERENCES

Aota, S., Gojobori, T., Ishibashi, F., Maruyama, T., and Ikemura, T. (1988). *Nucl. Acids Res.* **16**, r315–r402.

Grantham, R. (1980). *Trends Biochem. Sci.* **5**, 327–330.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A. (1980). *Nucl. Acids Res.* **8**, r49–r62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981). *Nucl. Acids Res.* **9**, r43–r74.

Ikemura, T. (1985). *Mol. Biol. Evol.* **2**, 13–24.

Ikemura, T., and Wata, K. (1991). *Nucl. Acids Res.* **19**, 4333–4339.

Maruyama, T., Gojobori, T., Aota, S., and Ikemura, T. (1986). *Nucl. Acids Res.* **14**, r151–r197.

Murray, E. E., Lotzer, J., and Eberle, M. (1989). *Nucl. Acids Res.* **17**, 477–494.

Wata, K., Aota, S., Tsuchiya, R., Ishibashi, F., Gojobori, T., and Ikemura, T. (1990). *Nucl. Acids Res.* **18**, r2367–r2411.