

Analysis of distribution of bases in the coding sequences by a diagrammatic technique

Chun-Ting Zhang and Ren Zhang¹

Department of Physics, Tianjin University, Tianjin 300072 and ¹School of Public Health, Shanghai Medical University, Shanghai 200032, China

Received June 21, 1991; Revised and Accepted September 11, 1991

ABSTRACT

The frequencies of occurrence of four bases in the first, second and third codon positions and in the total coding sequences have been calculated by the codon usage table published in 1990 by Ikemura *et al.* The distribution of frequencies are further analysed in detail by a graphic technique presented recently by us. Formulas expressing the frequencies of four bases in the first and second codon positions in terms of frequencies of amino acids have been given. It is shown by the graphic analysis that for 90 species, in the first codon position the purine bases are dominant and in most cases G is the most dominant base. In the second codon position A is the most dominant base, while G is the least dominant base. In the third codon position the G + C content varies from 0.1 to 0.9, keeping the A + C content equal to 1/2 and G content equal to that of C, approximately. If the frequencies for bases A, C, G and U in the total coding sequences are denoted by a , c , g and u , respectively, it is found that the unequal formula: $a^2 + c^2 + g^2 + u^2 < 1/3$, is valid for each of the 90 species including the human and *E.coli* etc.

INTRODUCTION

In their pioneering works in 1980 and 1981 in this journal, Grantham and his colleagues (1, 2) reported and analysed 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 11415 genes were compiled and analysed by Ikemura and his colleagues in 1986, 1988 and 1990 (3, 4, 5), respectively. Those of 207 higher plant genes were collected and analysed by Murray *et al.* in 1989 (6). To date, the codon usages reported by Ikemura *et al.* in 1990 (5) are the newest and largest set. A remarkable characteristic of the codon usage patterns has been pointed out by the above authors and other scientists. That is, the codon usage is nonrandom and species-specific. The codon usage patterns for different species are different. Among taxonomically related species the codon choice patterns resemble each other but they differ between distant species (5). Such a fact was regarded as a 'genome hypothesis' by Grantham *et al.* (1, 2). The same fact was called a 'codon dialect' by Ikemura (7). The situation is somewhat similar to the

light spectrum for an atom in physics. Each kind of atom has its own specific spectrum lines which are quite stable regardless of the number of atoms measured. In our case, there are '61 spectrum lines' for each organism. The different and stable distribution in magnitude of 61 spectrum lines represents different organism. It is possible to compare two organisms by comparing their corresponding spectrum lines. It is also possible to study the evolution of organisms by studying the evolution of codon usage patterns.

The codon usage tables are full of rich information. In our project of study we hope to find some new conclusions from the codon usage table as possible as we can. As a first step we shall study the distribution of DNA bases coding for the proteins by the codon usage table. The method that we use is a diagrammatic technique presented by us recently (8).

METHOD

There are four possible bases in the mRNA sequences coding for the proteins. Let the frequencies of occurrence of bases A, C, G and U be denoted by a , c , g and u , respectively. Obviously

$$a + c + g + u = 1, 0 < a, c, g, u < 1 \quad (1)$$

The above equation plays a key role in this study. As pointed out previously (8), we notice that the sum of distance of any point within a regular tetrahedron (RT) to the four faces is a constant, its height h . Setting the edge length of the RT equal to $\sqrt{6}/2$, then $h = 1$. Letting the four faces of the RT represent the four bases A, C, G and U, respectively, and letting the distance of a point P within the RT to the four faces A, C, G and U be equal to a , c , g and u , respectively, then we map the four real numbers a , c , g and u into a definite point P in this RT. For example, for $a = c = g = u = 1/4$, the mapping point P coincides just with the centre of the RT. The mapping is a one-to-one correspondence.

The values of a , c , g and u are easily calculated for each protein by the codon usage table when the data are available for that protein. In this case the mapping point in the RT represents that protein. Usually, the data for different proteins, but belonging to the same organism, are pooled. In this case, the values of a , c , g and u , and hence the mapping point in the RT represents that organism. Suppose that there are two organisms (or two

proteins) represented by two mapping points P_1 , (a_1, c_1, g_1, u_1) and P_2 , (a_2, c_2, g_2, u_2) , respectively. The distance between P_1 and P_2 was calculated as (8)

$$d_{12} = \frac{\sqrt{3}}{2} \left[(a_1 - a_2)^2 + (c_1 - c_2)^2 + (g_1 - g_2)^2 + (u_1 - u_2)^2 \right]^{1/2} \quad (2)$$

We define d_{12} as the 'distance' between the two organisms (or two proteins).

In order to study the distribution of the mapping points in a 3-dimensional space, it is convenient to project them onto some planes. Referring to Figure 1, consider an RT--BCGH. Let the regular triangle ΔBCG , ΔBGH , ΔBCH , ΔCGH represent the A-, C-, G- and U-face, respectively. The line connecting the middle point of an edge and that of the opposite edge is called the middle line of an RT. There are totally three middle lines in an RT, crossing in the centre O of the RT. They are perpendicular to each other. We can set up a Cartesian coordinate system OXYZ by using these three middle lines, as shown in Figure 1. The mapping points within this RT can be projected to any one of the three coordinate planes: X-Y, X-Z and Y-Z. Note that the projection of an RT to each coordinate plane is a square with side length $\sqrt{3}/2$, as shown in Figure 1. It is convenient to introduce the reduced coordinate system Oxyz, i.e.,

$$X = \frac{\sqrt{3}}{4} x, Y = \frac{\sqrt{3}}{4} y, Z = \frac{\sqrt{3}}{4} z, \quad (3)$$

then we obtain (8)

$$x = (a+g) - (u+c) = 2(a+g) - 1 = 1 - 2(u+c) \quad (4.a)$$

$$y = (a+c) - (u+g) = 2(a+c) - 1 = 1 - 2(u+g) \quad (4.b)$$

$$z = (a+u) - (g+c) = 2(a+u) - 1 = 1 - 2(g+c) \quad (4.c)$$

$$x, y, z \in [-1, 1]$$

or

$$\begin{bmatrix} a \\ c \\ g \\ u \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (5)$$

The distribution of mapping points can be studied by projecting them onto the x-y, x-z or y-z planes. As mentioned above, the projection of the RT to each of the x-y, x-z and y-z planes is a square with side length 2. Let us first pay attention to the x-y plane. Referring to Figure 2, the four vertexes of the square are called A-, C-, G- and U-vertex, respectively, for which we have

$$A: a = 1, c = g = u = 0; \quad G: g = 1, a = c = u = 0;$$

$$C: c = 1, a = g = u = 0; \quad U: u = 1, a = c = g = 0.$$

For the region $x > 0$, i.e., the region-AGKJ, $a+g > 1/2$, or the purines are dominant; $x < 0$, i.e., the region-JKUC, $c+u > 1/2$, or the pyrimidines dominant. Similarly, for $y > 0$, i.e., the region-AEFC, $a+c > 1/2$; for $y < 0$, i.e., the region-EGUF, $g+u > 1/2$. Consequently, in the first quadrant, $a+g > 1/2$ and $a+c > 1/2$; second, $c+u > 1/2$ and $a+c > 1/2$; third, $c+u > 1/2$ and $g+u > 1/2$; fourth, $a+g > 1/2$ and $g+u > 1/2$. Furthermore, for the two diagonals of the square, we have AOU: $g = c$; COG: $a = u$. Therefore, if the projective points are distributed within the region of ΔACG , it implies that $a > u$. Similarly, for ΔGCU , $u > a$; ΔAUG , $g > c$; ΔAUC , $c > g$. The square is divided by the two diagonals into four regions of triangles. For ΔAOG , called region-I, $a > u$ and $g > c$; for

ΔAOC , region-II, $a > u$ and $c > g$; for ΔCOG , region-III, $u > a$ and $c > g$; for ΔUOG , region-IV, $u > a$ and $g > c$. The above discussion is based on equation (4.a) and (4.b). We can have similar discussion about the x-z and y-z planes. However, the vertexes of square in the x-z plane should be associated with A, G, C and U (arranged in clockwise order), instead of A, G, U and C in the x-y plane. Similarly, the vertexes in the y-z plane are A, C, G and U, respectively. The detail is omitted here. The G+C content or $g+c$ is an important quantity in the study of nucleic acid sequences. Note that in the x-z or y-z planes, if the projecting points are situated at the region of $z > 0$, it implies that $g+c < 1/2$; otherwise, $z < 0$, i.e., $g+c > 1/2$.

In summary of the method, we first calculate the values of a, c, g and u for each kind of species by using the data of codon usage table (5). The four real numbers (a, c, g, u) are then mapped into a point within the RT. The points representing different species in the 3-dimensional space are projected onto the x-y, x-z and y-z planes, respectively, by using equations (4). The distance between any two points is calculated by equation (2). Then, some conclusions may be drawn by studying the distribution of the projecting points and the values of distance.

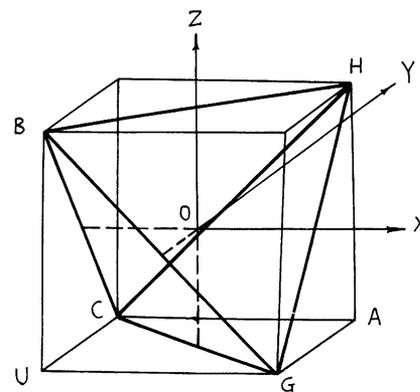


Figure 1. A cube and its inscribed regular tetrahedron.

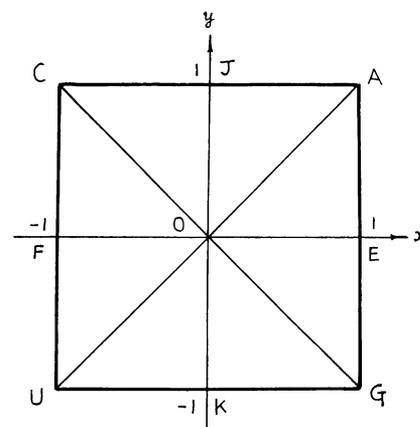


Figure 2. The projection of the regular tetrahedron to the x-y coordinate plane.

RESULTS AND DISCUSSION

The average codon usage data for 90 species or organelles were given by Ikemura *et al.* (5). In Table II of Ref. 5, for each species there are 61 values associated with the 61 codons. The frequencies of bases in the first, second and third position of codons are calculated by these data. The frequencies of four bases in the total coding sequences are also obtained. The distributions of bases are studied by the graphic method (Figures 3–5).

The distribution of frequencies of four bases in the first codon position for 90 species are shown in Figure 3a and b, where (a) means the projection to the x-y plane and (b) y-z plane. There are totally 90 points representing 90 species in each diagram. First, paying attention to Figure 3a and referring to Figure 2, we find the points are nearly all gathered in the region-I, i.e., $a > u$, $g > c$ or the region where the purine bases dominant. Since the points are situated near the x-axis, it implies $y \approx 0$ or $a + c \approx 1/2$, $g + u \approx 1/2$. Then look at Figure 3 (b). The points are near the z-axis, i.e., $y \approx 0$, the same conclusion as in Figure 3a. It is seen, for most points (about 7/9), $g + c > 1/2$. However,

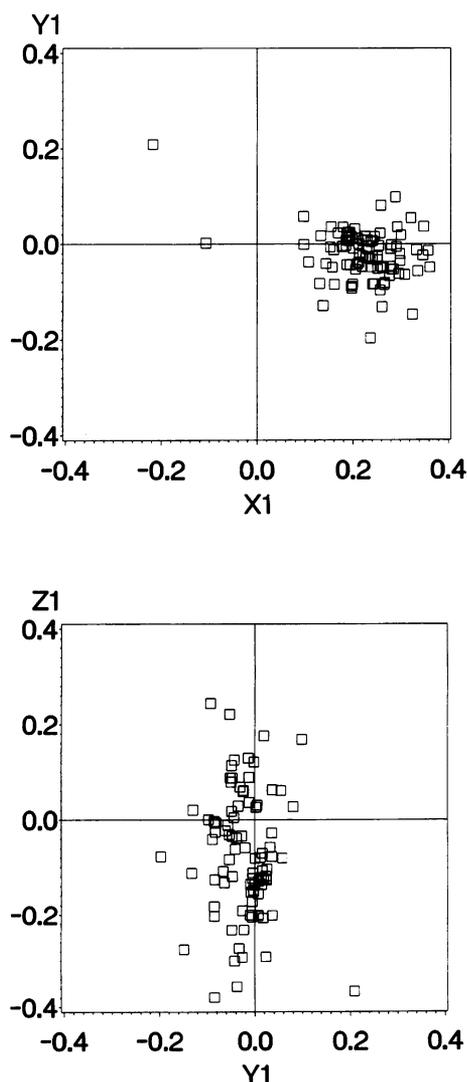


Figure 3. The distribution of frequencies of four bases in the first codon position for 90 species. (a), x-y plane and (b), y-z plane.

drawing two diagonals (not shown in Figure 3b), it is quite clear that for most points $g > a$ and $c > u$. Summarizing the results seen in Figure 3a and b, for nearly all of 90 species, we find $g > c$, $a > u$, $g + a > 1/2$, $a + c \approx 1/2$. For about 7/9 of 90 species, $g > a > c > u$. That is, in the first codon position the purine bases are dominant and in most cases G is the most dominant base.

In order to study the relations between the frequencies of bases in the first codon position and the frequencies of amino acids, we introduce three quantities α , β and γ , and define as

$$\alpha = [P(AGU) + P(AGC)]/P(\text{Ser}) \quad (6.a)$$

$$\beta = [P(UUA) + P(UUG)]/P(\text{Leu}) \quad (6.b)$$

$$\gamma = [P(AGA) + P(AGG)]/P(\text{Arg}) \quad (6.c)$$

where $P(AGU)$ means the frequency of codon AGU among 61 codons, and $P(\text{Ser})$ means the frequency of Ser among 20 amino acids, and so on. Obviously, $0 < \alpha, \beta, \gamma < 1$. Letting a_1 , c_1 , g_1 and u_1 represent the frequencies of bases A, C, G and U in the first codon position, respectively, then according to the genetic codes we have

$$a_1 = P(\text{Ile}) + P(\text{Met}) + P(\text{Thr}) + P(\text{Asn}) + P(\text{Lys}) + \alpha P(\text{Ser}) + \gamma P(\text{Arg}) \quad (7.a)$$

$$c_1 = (1 - \beta)P(\text{Leu}) + (1 - \gamma)P(\text{Arg}) + P(\text{Pro}) + P(\text{His}) + P(\text{Gln}) \quad (7.b)$$

$$g_1 = P(\text{Val}) + P(\text{Ala}) + P(\text{Gly}) + P(\text{Glu}) + P(\text{Asp}) \quad (7.c)$$

$$u_1 = \beta P(\text{Leu}) + (1 - \alpha)P(\text{Ser}) + P(\text{Phe}) + P(\text{Tyr}) + P(\text{Cys}) + P(\text{Trp}) \quad (7.d)$$

The above conclusion that $g_1 > c_1$, $a_1 > u_1$, $g_1 + a_1 > 1/2$, $a_1 + c_1 \approx 1/2$, must exert some restriction on the frequencies of amino acids occurring in proteins in these 90 species. For example, simple mathematical consideration shows that $0.3 \leq g_1 < 0.5$, this implies

$$0.3 \leq P(\text{Val}) + P(\text{Ala}) + P(\text{Gly}) + P(\text{Glu}) + P(\text{Asp}) < 0.5 \quad (8)$$

Since the points in Figure 3 are distributed within a considerably small area (note that the scales of the coordinate are small as compared with Figure 5b), we conclude that the divergence of amino acid frequencies for different species is small, too. In other words, the distributions of frequencies of amino acids for 90 species are similar. This conclusion has been confirmed by a direct calculation of the amino acid frequencies for each of the 90 species (data not shown here for saving the space).

The distribution of frequencies of the four bases in the second codon position for 90 species are shown in Figure 4a and b, respectively, for the x-y and y-z projection planes. First look at Figure 4a and refer to Figure 2. The points are nearly all distributed in the region $y > 0$, i.e., $a + c > 1/2$ or $g + u < 1/2$. At the same time, the points are in the region-II and III, i.e., $c > g$. Next turn to Figure 4b. The points are nearly all in the first quadrant, i.e., $y > 0$, $z > 0$, or $a + c > 1/2$, $a + u > 1/2$. At the same time, the points are nearly all in the region-II and gather near by the diagonal $c = u$, so $a > g$, $c \geq u$. Summarizing the results seen in Figure 3a and b, we find $a > c \geq u > g$, $a + c > 1/2$, $a + u > 1/2$. That is, in the second codon position A is the most dominant base, while G is the least dominant base. Similar to equation (7), we have

$$a_2 = P(\text{Tyr}) + P(\text{His}) + P(\text{Gln}) + P(\text{Asp}) + P(\text{Lys}) + P(\text{Asn}) + P(\text{Glu}) \quad (9.a)$$

$$c_2 = (1 - \alpha)P(\text{Ser}) + P(\text{Pro}) + P(\text{Thr}) + P(\text{Ala}) \quad (9.b)$$

$$g_2 = P(\text{Arg}) + P(\text{Cys}) + P(\text{Trp}) + P(\text{Gly}) + \alpha P(\text{Ser}) \quad (9.c)$$

$$u_2 = P(\text{Phe}) + P(\text{Leu}) + P(\text{Ile}) + P(\text{Met}) + P(\text{Val}) \quad (9.d)$$

There are also some restrictions on the frequencies of amino acids. For example, since $c_2 > g_2$, it implies

$$(1 - 2\alpha)P(\text{Ser}) + P(\text{Pro}) + P(\text{Thr}) + P(\text{Ala}) > P(\text{Arg}) + P(\text{Cys}) + P(\text{Trp}) + P(\text{Gly}). \quad (10)$$

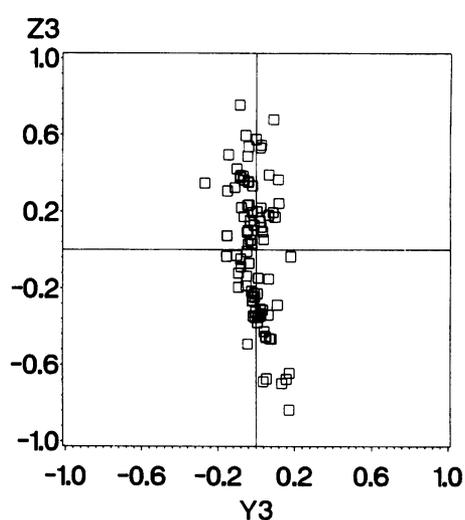
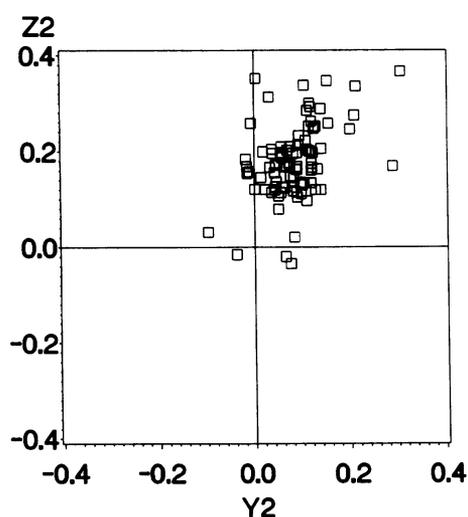
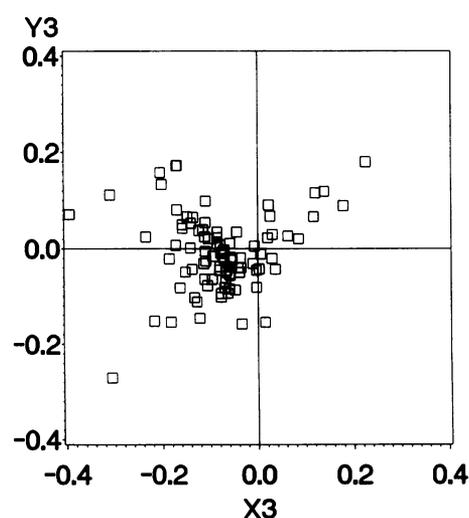
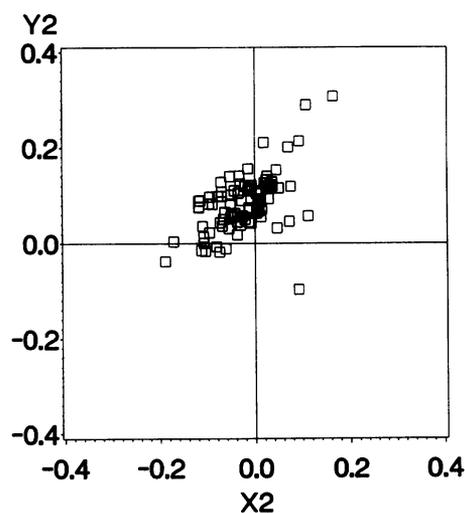


Figure 4. The distribution of frequencies of four bases in the second codon position for 90 species. (a), x-y plane and (b), y-z plane.

Figure 5. The distribution of frequencies of four bases in the third codon position for 90 species. (a), x-y plane and (b), y-z plane. Note that the scale of Figure 5b is larger than those of Figures 3, 4 and 5a.

The fact that $g_1 > g_2$ was called a G-non-G periodicity in the coding sequences by Trifonov (9). We define a quantity σ as

$$\sigma = g_1/g_2 \quad (11)$$

which may be useful in the study of G-non-G periodicity in the coding sequences. The typical value of σ is nearly equal to 2.

The distribution of frequencies of the four bases in the third codon position for 90 species are shown in Figure 5a and b, for the x-y and y-z projection planes, respectively. At first, look at Figure 5a. The points are in the region $y \approx 0$, $x \leq 0$, i.e., $a + c \approx 1/2$, $a + g \leq 1/2$, and for most points $g \approx c$. From Figure 5b we find $y \approx 0$, the same as seen in Figure 5a. However, the values of z vary in a large interval, i.e., from -0.8 to $+0.8$, or $g + c$ from 0.9 to 0.1 . Summarizing the results seen in Figure 5a and b, we find the values of $g + c$ vary from 0.1 to 0.9 , keeping $a + c \approx 1/2$, $a + g \leq 1/2$ and $g \approx c$. By counting the number of points we find the number of points in the region $z > 0$ (or $g + c < 1/2$) is equal to that in $z < 0$ (or $g + c > 1/2$). The situation is quite different with that in the first and second codon position. In the case of third position of codon we cannot find equations

similar to (7) or (9). Therefore, the frequencies of amino acids have nothing to do with the values of a_3 , c_3 , g_3 and u_3 .

By using the formula $b = (b_1 + b_2 + b_3)/3$, where b represents a , c , g and u , respectively, the distribution of frequencies of the four bases in the total coding sequences for 90 species are analysed by a similar technique of diagram (not shown here for saving the space). The inscribed sphere of the RT was ever called the sphere of nucleic acid (SNA) (8). The projection of SNA to each coordinate plane is a circle with radius of $1/\sqrt{3} \approx 0.577$. An empirical formula was found for the genomic DNA that the mapping points are nearly all distributed within the SNA for different species, or equivalent to say (8)

$$1/4 \leq a^2 + c^2 + g^2 + u^2 < 1/3 \quad (12)$$

We have found that equation (12) is completely satisfied for each of the 90 species. That is, the frequencies in the total coding sequences for each of the 90 species obey the requirement that $s = a^2 + c^2 + g^2 + u^2 < 1/3$. The reason why $s < 1/3$ is still not clear.

CONCLUSION

The distribution of frequencies of the four bases for the coding sequences of 90 species (5) have been studied by a graphic technique. The overall characteristics of distribution of frequencies of four bases in the first, second and third codon positions have been discussed in detail for 90 species. It is shown that in the first position the purine bases are dominant and in most cases G is the most dominant base. In the second position A is the most dominant base, while G is the least dominant base. In the third position $g+c$ varies from 0.1 to 0.9, keeping $a+g \approx 1/2$ and $g \approx c$. Formulas connecting the frequencies of the four bases in the first and second codon positions with the frequencies of amino acids have been presented. The fact that the distributions of mapping points gather in a small area for the first and second codon positions shows that the distributions of frequencies of amino acids for different species are similar. The frequencies of four bases in the total coding sequences have been also calculated. It is shown that the sum of square of frequencies of four bases is less than $1/3$ for each of the 90 species.

The diagrammatic technique that we use may be a useful tool in the study of distribution of bases. It does help to summarize a lot of data in a readily perceivable form. In addition, it is quite easy to use. As long as the frequencies of bases are known, one can easily draw the diagram by using equation (4). Generally speaking, the users need not concerning the principle of the diagram based on the geometry of regular tetrahedron. Therefore, we hope that the diagrammatic technique will become an auxiliary and intuitional tool for the researchers in the area of nucleic acid research.

REFERENCES

1. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) *Nucl. Acids Res.* **9**, r49–r62.
2. Grantham,R., Gautier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981) *Nucl. Acids Res.* **9**, r43–r74.
3. Maruyama,T., Gojobori,T., Aota,S. and Ikemura,T. (1986) *Nucl. Acids Res.* **14**, r151–r197.
4. Aota,S., Gojobori,T., Ishibashi,F., Maruyama,T. and Ikemura,T. (1988) *Nucl. Acids Res.* **16**, r315–r402.
5. Wata,K., Aota,S., Tsuchiya,R., Ishibashi,F., Gojobori,T. and Ikemura,T. (1990) *Nucl. Acids Res.* **18**, r2367–r2411.
6. Murray,E.E., Lotzer,J. and Eberle,M. (1989) *Nucl. Acids Res.* **17**, 477–494.
7. Ikemura,T. (1985) *Mol. Biol. Evol.* **2**, 13–24.
8. Zhang,C.-T. and Zhang,R. (1991) *Int. J. Biol. Macromol.* **13**, 45–49.
9. Trifonov,E.N. (1987) *J. Mol. Biol.* **194**, 643–652.