Amino Acids (1997) 12: 167–177

Amino
Acids

Distribution of mapping points of 20 amino acids in the tetrahedral space

R. Zhang

Department of Epidemiology and Biostatistics, Tianjin Cancer Hospital and Institute, Tianjin, China

Accepted July 26, 1996

Summary. Based on the genetic codes and a simple theorem for the geometrical property of the regular tetrahedron, each amino acid is mapped onto a unique point in a 3-dimensional tetrahedral space. The distribution of the 20 mapping points for 20 amino acids is studied in detail. It is found that the mapping points for the hydrophobic and hydrophilic amino acids are distributed at distinct regions in the 3-dimensional space. A plane separating the two kinds of points satisfactorily based on the Fisher's algorithm has been calculated. It is shown that the codons coding for the hydrophobic amino acids are constituted dominantly by the bases of keto group, i.e., G and T. While the codons coding for the hydrophilic amino acids are constituted dominantly by the bases of amino group, i.e., A and C. The biological implication of the mapping points and the separating plane has been discussed in some details.

Keywords: Amino acids – Mapping point – Hydrophobicity – Hydrophilicity – Separating plane – Fisher's algorithm

Introduction

The physical and chemical characteristics of amino acids are determined by their chemical structures. On the other hand, the amino acids are encoded by the genetic codons. Therefore, the physical and chemical characteristics of amino acids should be relevant to the genetic codons. Among the characteristics of amino acids the hydrophobicity and hydrophilicity seem to be the most important factors determining the folding structures of proteins. In this paper we would like to explore the relationship between the hydrophobic – hydrophilic characteristics of amino acids and the codons coding for them.

The method that we use in this study is based on our previous work (Zhang and Zhang, 1991a,b; 1994). The base composition for any given DNA sequence can be calculated easily. The base composition for a given DNA sequence is then mapped onto a unique point within a tetrahedral space. Since any codon may be regarded as a DNA sequence with three bases, there exists

a unique mapping point within the tetrahedron, corresponding to the codon concerned. In case of degenerate, the centroid of the mapping points associated with degenerate codons is regarded as the mapping point for the amino acid encoded by the degenerate codons. In case of no-degenerate, the mapping point itself for the codon is regarded as the mapping point for the amino acid encoded by this codon. In such a way, there are uniquely 20 points in a three-dimensional space, corresponding to each of the 20 amino acids, respectively. Then we study the distribution of the 20 points in the three-dimensional space. We hope to see if they can be grouped into two catalogues, each of which is associated with the hydrophobic or hydrophilic amino acids. It should be pointed out that a similar representation for the 64 codons by a tetrahedron was proposed by Trainor et al. (1984). However, the representation points for the 64 codons are situated at the four faces of the tetrahedron in Trainor's approach. While in our method, the 20 mapping points representing the 20 amino acids are situated at the interior of the tetrahedron. Our representation provides the possibility for further studying the relationship between the characteristics of amino acids and the genetic codes by a geometrical approach. Therefore, the two representations are distinct from each other essentially.

Method

Representation of 20 amino acids by 20 mapping points in a 3-dimensional space

Suppose that the frequencies of occurrence of the base A, C, G and T in a single-strand DNA sequence are denoted by a, c, g and t, respectively. Obviously,

$$a + c + g + t = 1. \tag{1}$$

Equation (1) plays a key role in this study. This equation describes the constraint condition applied to the four real numbers a, c, g and t. On the other hand, there exists a simple theorem about a regular tetrahedron, which may be expressed as follows: *The sum of the four distances to the four faces from any point within a regular tetrahedron must be equal to a constant, its height.* Now imagine a special regular tetrahedron, whose height is equal to 1. If the four real numbers a, c, g and t are associated with the four distances described above, the point within the tetrahedron constitutes a mapping of one-to-one correspondence for the four numbers. The three middle lines of the regular tetrahedron, crossing in its center, are perpendicular to each other. The Cartesian coordinate system is set up by using the three middle lines, as shown in Fig. 1. The coordinates x, y and z of the mapping point associated with the four real numbers may be expressed as follows (Zhang and Zhang, 1991a,b)

$$\begin{cases} x = 2(a + g) - 1, \\ y = 2(a + c) - 1, \\ z = 2(a + t) - 1. \end{cases} \tag{2}$$

Starting from the genetic codes, we can find out the mapping point for each amino acid. The procedure may be described as follows. Consider the amino acid Trp first. The codon coding for Trp is TGG. Accordingly, we find $a = 0$, $c = 0$, $g = 2/3$, and $t = 1/3$. Substituting these numbers into eqs. (2), we find $x = 1/3$, $y = -1$, and $z = -1/3$. In case

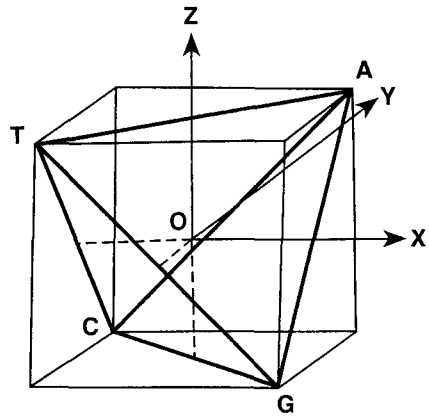

Fig. 1. A cube and its inscribed regular tetrahedron. Note that the three middle lines of the regular tetrahedron, crossing in its centre, are perpendicular to each other. The coordinate system is set up by using the three middle lines. The four bases *A*, *C*, *G* and *T* are assigned to the four vertices of the tetrahedron, respectively

Table 1. Coordinates of mapping points for 20 amino acids[a]

A.A.[b]	x_s	y_s	z_s	A.A.[b]	x_q	y_q	z_q
Ile	−1/9	1/9	7/9	Pro	−2/3	2/3	−2/3
Phe	−1	−2/3	2/3	Thr	0	2/3	0
Val	0	−2/3	0	Ser	−1/3	0	0
Leu	−5/9	−2/9	2/9	His	−1/3	2/3	0
Trp	1/3	−1	−1/3	Glu	1	0	0
Met	1/3	−1/3	1/3	Asn	1/3	2/3	2/3
Ala	0	0	−2/3	Gln	1/3	2/3	0
Gly	2/3	−2/3	−2/3	Asp	1/3	0	0
Cys	−1/3	−2/3	0	Lys	1	2/3	2/3
Tyr	−1/3	0	2/3	Arg	1/3	0	−4/9

[a] The subscripts "s" and "q" correspond to the hydrophobic and hydrophilic amino acids, respectively. [b] The order of amino acids is by decreasing hydrophobicity on the consensus scale of Eisenberg (1984).

of degenerate, as described above, the centroid of the mapping points associated with all of the degenerate codons is used to represent the amino acid concerned. For example, Phe is encoded by two codons TTT and TTC. The average coordinates of the two mapping points are $x = -1$, $y = -2/3$, and $z = 2/3$. Therefore, this point is used to represent the amino acid Phe. The final coordinates of 20 mapping points for each of the 20 amino acids are listed in Table 1.

The boundary plane between the hydrophobic and hydrophilic mapping points

The 20 amino acids may be classified into two categories i.e., the hydrophobic and hydrophilic ones. However, as pointed out by Eisenberg (1984), no generally accepted method exists to measure or calculate hydrophobicities, each method used constitutes a separate operational definition. The consensus scale of hydrophobicity defined by Eisenberg (1984) was designed to mitigate the effects of outlying values in any one scale,

produced by the peculiarities of the method (Eisenberg, 1984). The consensus scales of Eisenberg are used in this study. The order of 20 amino acids by decreasing hydrophobicity on the consensus scale is listed as follows: Ile, Phe, Val, Leu, Trp, Met, Ala, Gly, Cys, Tyr, Pro, Thr, Ser, His, Glu, Asn, Gln, Asp, Lys and Arg. The scales of the former ten amino acids are positive, thus which are regarded as hydrophobic ones; the scales of the latter ten amino acids are negative, thus which are regarded as hydrophilic ones. As pointed out above, each amino acid is represented by a unique mapping points in a 3-dimensional space. For convenience, the 10 mapping points for the hydrophobic amino acids are always denoted by open circles, while the 10 mapping points for the hydrophilic amino acids are denoted by filled circles, in all the relevant figures in this paper.

Our task is to find out a plane in the 3-dimensional space, which separates the two kinds of points best. What do we mean by "best"? Generally speaking, there is no answer for this question. Because the answer depends on the criterion used in the calculation. Here we use the Fisher's algorithm to calculate the best separating plane (see, e.g., Anderson, 1984). Because the separating criterion by the Fisher's algorithm is recognized as considerably reasonable.

Suppose that the centriods of the mapping points for the hydrophobic and hydrophilic amino acids are denoted by P_s and P_q, respectively. The coordinates of P_s are denoted by $\bar{x}_s$, $\bar{y}_s$ and $\bar{z}_s$. The coordinates of P_q are denoted by $\bar{x}_q$, $\bar{y}_q$ and $\bar{z}_q$. We have

$$\bar{x}_s = \frac{1}{10}\sum_{i=1}^{10} x_{si}, \qquad \bar{y}_s = \frac{1}{10}\sum_{i=1}^{10} y_{si}, \qquad \bar{z}_s = \frac{1}{10}\sum_{i=1}^{10} z_{si}, \tag{3}$$

$$\bar{x}_q = \frac{1}{10}\sum_{i=1}^{10} x_{qi}, \qquad \bar{y}_q = \frac{1}{10}\sum_{i=1}^{10} y_{qi}, \qquad \bar{z}_q = \frac{1}{10}\sum_{i=1}^{10} z_{qi}, \tag{4}$$

where x_{si}, y_{si} and z_{si} are the coordinates of the mapping point of the *ith* hydrophobic amino acid in the order listed in Table 1. For example, x_{s1}, y_{s1} and z_{s1} are the coordinates for Ile, and so forth. The meaning of x_{qi}, y_{qi} and z_{qi} is quite similar. For example, x_{q1}, y_{q1} and z_{q1} are the coordinates for Pro, and so forth.

Suppose further that the best separating plane in the sense of Fisher's algorithm is expressed by the following equation

$$\text{S:} \qquad c_0 + c_1 x + c_2 y + c_3 z = 0, \tag{5}$$

where c_i $(i = 0, 1, 2, 3)$ are the plane parameters to be determined later. These parameters can be determined uniquely by the following three rules.

Rule 1. Project the hydrophobic and hydrophilic centroids as defined in eqs. (3), (4), respectively, onto the normal direction of the desired plane. The distance between these two projected points should be kept as large as possible.

Rule 2. Along the normal direction of the plane, the sum of the squares of the distances between the hydrophobic centroid and each of the mapping points of the hydrophobic amino acids should be as small as possible. The same is true for the sum of squares of the distances between the hydrophilic centroid and each of the mapping points of the hydrophilic amino acids. In other words, the sum of the two sums described above should be kept as small as possible.

Rule 3. The desired plane should pass through the average position of the hydrophobic and hydrophilic centroids.

In order to determine the parameters c_i $(i = 0, 1, 2, 3)$ so as to uniquely fix the desired plane S, all the above three rules must be satisfied. According to the analytic geometry, the square of the projected distance between the hydrophobic and hydrophilic centroids along the normal direction of the plane S, denoted by D^2, is

$$D^2 = \frac{\left[c_1(\bar{x}_s - \bar{x}_q) + c_2(\bar{y}_s - \bar{y}_q) + c_3(\bar{z}_s - \bar{z}_q)\right]^2}{c_1^2 + c_2^2 + c_3^2} = \frac{\tilde{D}^2}{c_1^2 + c_2^2 + c_3^2}, \tag{6}$$

where

$$\tilde{D}^2 = \left[c_1(\bar{x}_s - \bar{x}_q) + c_2(\bar{y}_s - \bar{y}_q) + c_3(\bar{z}_s - \bar{z}_q)\right]^2. \tag{7}$$

Similarly, along the normal direction of the plane, the sum of the squares of the projected distances between the hydrophobic centroid and each of the mapping points of the hydrophobic amino acids may be expressed as

$$D_s^2 = \frac{\tilde{D}_s^2}{c_1^2 + c_2^2 + c_3^2}, \tag{8}$$

where

$$\tilde{D}_s^2 = \sum_{i=1}^{10}\left[c_1(x_{si} - \bar{x}_s) + c_2(y_{si} - \bar{y}_s) + c_3(z_{si} - \bar{z}_s)\right]^2. \tag{9}$$

Similarly, the sum of the squares of the projected distances along the normal direction of the plane between the hydrophilic centroid and each of the mapping points of the hydrophilic amino acids may be expressed as

$$D_q^2 = \frac{\tilde{D}_q^2}{c_1^2 + c_2^2 + c_3^2}, \tag{10}$$

where

$$\tilde{D}_q^2 = \sum_{i=1}^{10}\left[c_1(x_{qi} - \bar{x}_q) + c_2(y_{qi} - \bar{y}_q) + c_3(z_{qi} - \bar{z}_q)\right]^2. \tag{11}$$

The rule 1 requires the value of D^2 kept as large as possible; while the rule 2 requires the sum of D_s^2 and D_q^2 kept as small as possible. In order to find the desired plane, we define the objective function O as follows

$$O(c_1, c_2, c_3) = \frac{D^2}{D_s^2 + D_q^2} = \frac{\tilde{D}^2}{\tilde{D}_s^2 + \tilde{D}_q^2} = \frac{\tilde{D}^2}{\tilde{F}^2}, \tag{12}$$

where

$$\tilde{F}^2 = \tilde{D}_s^2 + \tilde{D}_q^2. \tag{13}$$

The rule 1 and rule 2 require that the objective function O should be kept as large as possible. In other words, the plane parameters c_1, c_2 and c_3 should be determined by the way such that the objective function O reaches its maximum. Accordingly, we have

$$\frac{\partial O}{\partial c_i} = 0, \qquad (i = 1, 2, 3). \tag{14}$$

Or

$$\frac{\partial O}{\partial c_i} = \frac{\tilde{F}^2 \partial \tilde{D}^2/\partial c_i - \tilde{D}^2 \partial \tilde{F}^2/\partial c_i}{\tilde{F}^4} = 0. \tag{15}$$

Substituting eq. (12) into eq. (15), we find

$$\frac{1}{O}\frac{\partial \tilde{D}^2}{\partial c_i} = \frac{\partial \tilde{F}^2}{\partial c_i}, \qquad (i = 1, 2, 3). \tag{16}$$

Omitting some trivial derivation steps, we obtain the following linear algebraic equations for c_1, c_2, c_3

$$\begin{cases} T_{xx}c_1 + T_{xy}c_2 + T_{xz}c_3 = d_1, \\ T_{yx}c_1 + T_{yy}c_2 + T_{yz}c_3 = d_2, \\ T_{zx}c_1 + T_{zy}c_2 + T_{zz}c_3 = d_3, \end{cases} \tag{17}$$

where

$$\begin{cases} d_1 = \bar{x}_s - \bar{x}_q, \\ d_2 = \bar{y}_s - \bar{y}_s, \\ d_3 = \bar{z}_s - \bar{z}_q, \end{cases} \tag{18}$$

and

$$T_{mn} = \sum_{i=1}^{10}(m_{si} - \bar{m}_s)(n_{si} - \bar{n}_s) + \sum_{i=1}^{10}(m_{qi} - \bar{m}_q)(n_{qi} - \bar{n}_q), \tag{19}$$

$$m, n = x, y, z. \tag{20}$$

Note that

$$T_{mn} = T_{nm}. \tag{21}$$

As soon as the solution of eqs. (17) is found, the value of c_0 can be calculated by the rule 3

$$c_0 = -c_1x_0 - c_2y_0 - c_3z_0, \tag{22}$$

where

$$\begin{cases} x_0 = \frac{1}{2}(\bar{x}_s + \bar{x}_q), \\ y_0 = \frac{1}{2}(\bar{y}_s + \bar{y}_q), \\ z_0 = \frac{1}{2}(\bar{z}_s + \bar{z}_q). \end{cases} \tag{23}$$

In fact, x_0, y_0 and z_0 are the coordinates of the average position of the hydrophobic and hydrophilic centroids, which is the overall centroid of the whole mapping points.

Result and discussion

The distribution of the 20 mapping points for 20 amino acids are shown in Fig. 2, where the hydrophobic points are denoted by open circles; while the hydrophilic points are denoted by filled circles, as described above. As can be seen that the two kinds of points are distributed roughly at different regions in the 3-dimensional space. This result reflects the fact that the hydrophobicity

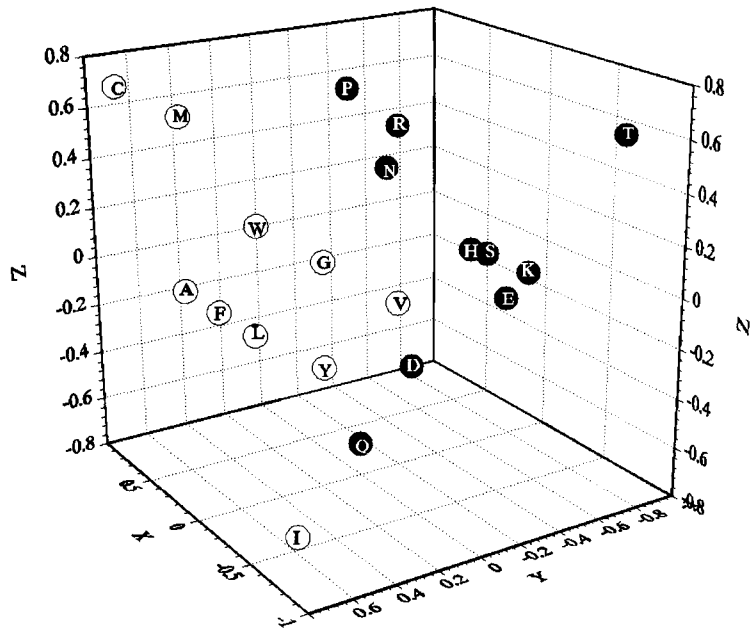

Fig. 2. The distribution of the 20 mapping points for 20 amino acids in the 3-dimensional tetrahedral space. Each amino acid is denoted by its single letter code. The hydrophobic (hydrophilic) amino acids are represented by open (filled) circles

and hydrophilicity of amino acids are strongly determined by the base composition of the codons coding for the two kinds of amino acids. Our approach has the merit that the characteristic pattern is easily detected visually by the researchers. It helps the researchers in summarizing the data in a readily perceivable form. Furthermore, the method adopted in this study leads to a new research area that the problem is studied by a pure geometrical approach. The two kinds of mapping points are separated by the best separating plane in the sense of Fisher's algorithm. This plane is described by eq. (4), where the plane parameters are uniquely determined by the Fisher's algorithm, as described in details in the Method section. The result is listed as follows

$$\begin{cases} c_0 = 0.018, \\ c_1 = 0.130, \\ c_2 = 0.325, \\ c_0 = -0.130. \end{cases} \tag{24}$$

To display the result more clearly, it is of convenience to project the mapping points onto some coordinate planes. Here the x-y and y-z coordinate planes are used. The line of intersection between the separating plane and the x-y plane is described by the equation

$$c_0 + c_1 x + c_2 y = 0. \tag{25}$$

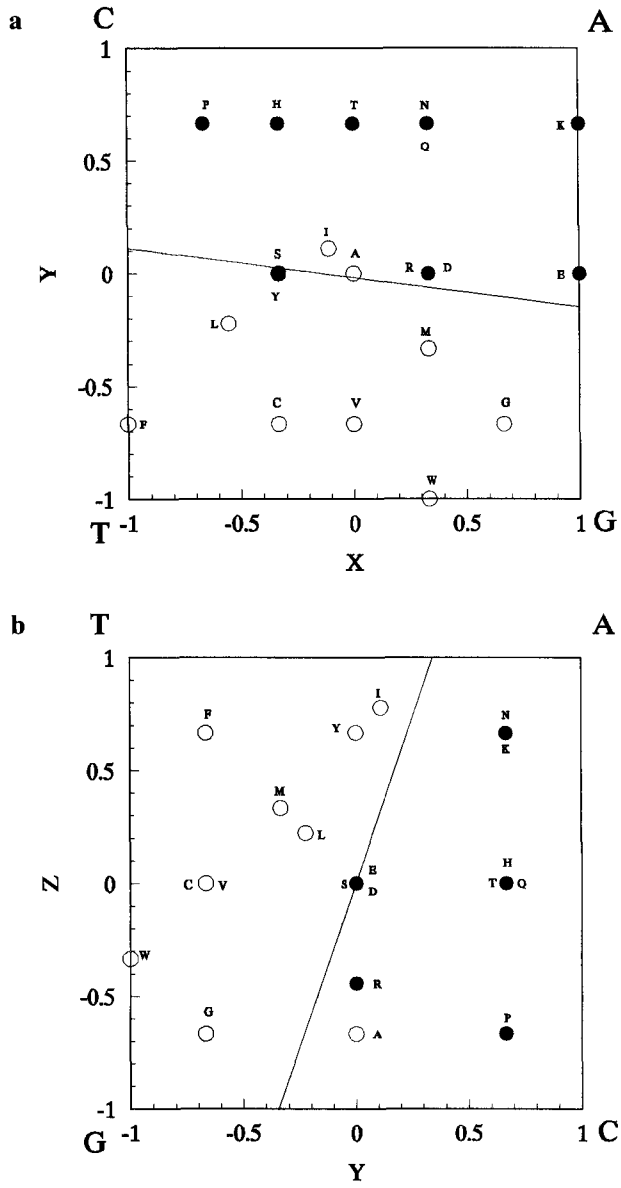

Fig. 3. The projection of the 20 mapping points for 20 amino acids onto **a** the x-y coordinate plane and **b** the y-z plane. Each amino acid is denoted by its single letter code. The hydrophobic (hydrophilic) amino acids are represented by open (filled) circles. The line in **a** (**b**) is the line of intersection between the separating plane and the x-y (y-z) plane. Note that the projection of the tetrahedron onto any coordinate plane, set up as shown in Fig. 1, is just a regular square. The four letters A, C, G and T near by the four vertices of the square represent the bases of DNA. For more details, see Zhang and Zhang (1991a,b)

The line of intersection between the separating plane and the y-z plane is described by the equation

$$c_0 + c_2 y + c_3 z = 0. \tag{26}$$

The projections of the mapping points onto the x-y and y-z coordinate planes and the lines of intersection are shown in Fig. 3a and 3b, respectively. The projected point for each amino acid is represented by its single letter code. Summarizing the overall result seen in Fig. 3a and b, we have found that the best separating plane separates the two kinds of mapping points satisfactorily. The only exception is alanine (Ala). According to the consensus scale of Eisenberg (1984), Ala is of hydrophobic amino acid. Nevertheless, its mapping points is situated at the side of those of hydrophilic amino acids. Another two amino acids Ser and Lys are situated almost at the separating plane itself. The existence of such a separating plane implies that the mapping points of 20 amino acids based on their genetic codes are distributed naturally at two distinct regions. This is an interesting finding of the present study.

We discuss further the biological meaning of the above result. The coordinates of the hydrophobic and hydrophilic centroids are found as

$$\begin{cases} \bar{x}_s = -0.100, \\ \bar{y}_s = -0.411, \\ \bar{z}_s = 0.100. \end{cases} \tag{27}$$

$$\begin{cases} \bar{x}_q = 0.167, \\ \bar{y}_q = 0.267, \\ \bar{z}_q = -0.111. \end{cases} \tag{28}$$

Eq. (2) may be written as

$$\begin{pmatrix} a \\ c \\ g \\ t \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{29}$$

Substituting eqs. (27) and (28) into eq. (29), we obtain the base compositions corresponding to the hydrophobic and hydrophilic centroids, respectively

$$\begin{cases} \bar{a}_s = 0.147, \\ \bar{c}_s = 0.147, \\ \bar{g}_s = 0.303, \\ \bar{t}_s = 0.403, \end{cases} \tag{30}$$

$$\begin{cases} \bar{a}_q = 0.331, \\ \bar{c}_q = 0.303, \\ \bar{g}_q = 0.253, \\ \bar{t}_q = 0.113, \end{cases} \tag{31}$$

where the subscripts "s" and "q" are associated with the hydrophobic and hydrophilic amino acids, respectively. This result implies that the hydrophobic amino acids are encoded by the codons mainly constituted by the bases of keto group (G + T). While the hydrophilic amino acids are encoded by the codons mainly constituted by the bases of amino group (A + C). This result can be seen directly by observing Fig. 3a and b. Note that the line of intersection between the separating plane and the x-y plane coincides approximately with the line of $y = 0$. In other words, the projected points of the mapping points for the hydrophilic amino acids are situated roughly at the region of $y > 0$ in Fig. 3a. According to the principle of the diagrammatic technique (Zhang and Zhang, 1991a,b), the bases of amino group are predominant in this case. At the same time, the bases of keto group are predominant for the hydrophobic case. The same conclusion can be also obtained by observing Fig. 3b. It is very interesting to point out that the separating plane passes through the center of the tetrahedron approximately, due to the fact that $c_0 \approx 0$. Therefore, the tetrahedron is divided by the separating plane into two equal parts roughly.

Substituting eqs. (2) and (24) into eq. (5), we obtain

$$a_p + c_p = 0.472 + 0.4 \times (t_p - g_p), \tag{32}$$

where a_p, c_p, g_p and t_p are the base composition associated with the mapping points situated at the separating plane S. For the mapping points situated at the side of the hydrophobic amino acids, we have

$$a + c < 0.472 + 0.4 \times (t - g). \tag{33}$$

Similarly, for the mapping points situated at the side of the hydrophilic amino acids, we have

$$a + c > 0.472 + 0.4 \times (t - g). \tag{34}$$

Eqs. (33) and (34) show the characteristics of the base compositions for the codons coding for the hydrophobic and hydrophilic amino acids, respectively.

The methodology presented in this paper is considerably novel. This approach has the merit that the characteristic distribution pattern of the mapping points is easily detected visually by the researchers. It helps the researchers in summarizing the data in a readily perceivable form. Furthermore, the method adopted in this study leads to a new research area that the problem is solved by a pure geometrical approach. As pointed out by Eisenberg (1984) that no generally accepted method exists to measure or calculate hydrophobicities. Our approach provides a new way to study the problem from a geometrical point of view. Further investigation is being underway in our laboratory.

Acknowledgement

The author is grateful to Professor Qing-Sheng Wang for his support during this study.

References

Anderson TW (1984) An introduction to multivariate statistical analysis. 2nd edn. Wiley, New York

Eisenberg D (1984) Three-dimensional structure of membrane and surface proteins. Ann Rev Biochem 53: 595–623

Trainor LEH, Rowe GW, Szabo VL (1984) A tetrahedral representation of poly-codon sequences and a possible origin of codon degeneracy. J Theor Biol 108: 459–468

Zhang CT, Zhang R (1991a) Diagrammatic representation of the distribution of DNA bases and its applications. Int J Biol Macromol 13: 45–49

Zhang CT, Zhang R (1991b) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucleic Acids Res 19: 6313–6317

Zhang R, Zhang CT (1994) Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. J Biomol Struc Dynamics 11: 767–782

Author's address: Ren Zhang, M.D., Department of Epidemiology and Biostatistics, Tianjin Cancer Hospital and Institute, Tianjin 300060, China.

Received February 28, 1996