

Environment-dependent and position-specific frequencies of amino acid occurrences in α -helices

Hiroshi Wako^{1*}, Jianghong An² and Akinori Sarai²

¹*School of Social Sciences, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan*

²*Tsukuba Institute, The Institute of Physical & Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan*

*E-mail: wako@waseda.jp

(Received March 31, 2003; accepted May 14, 2003; published online May 31, 2003)

Abstract

By using the method to define a local structural motif of proteins by the Delaunay tessellation proposed by Wako and Yamato (*Protein Eng.* 11, 981-990 (1998)), we analyzed environment-dependent and position-specific frequencies of amino-acid occurrences in α -helices. In that method the three-dimensional structure of a protein molecule is uniquely divided into non-overlapping Delaunay tetrahedrons, each vertex of which is occupied by one of the comprising residues. A code is then assigned to each tetrahedron so as to characterize the local structure containing it. The tetrahedrons located in the interior of the α -helices are assigned 36 kinds of codes. The differences in the codes reflect the existence and absence of four residues surrounding the relevant region of the α -helix. In other words, the environment of the α -helix can be differentiated by these codes. Accordingly, we analyzed the frequencies of amino acid occurrences on each vertex of the tetrahedrons for each of these codes. Such data provide information about possible amino acid substitutions specific to a vertex position (*i.e.*, a position in the α -helix) for a given code (*i.e.*, environment around the α -helix). Furthermore, the principal component analysis was carried out to reveal general features of the amino acid occurrences in the α -helices. In relation to these results, such frequencies at the N- and C-terminals of the α -helix are also discussed.

Key Words: Delaunay tessellation, local structure motif, amino acid substitution, principal component analysis, helix capping

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

An α -helix is the most easily recognizable local structure in protein structures owing to its regularity, and an important structural element for protein folding. Many efforts have been made to

predict the locations of the α -helices in a protein from its amino acid sequence. For such a purpose, many experimental and theoretical data have been provided for determination of the propensities of amino acid residues to occur in α -helices [1][2][3][4][5][6][7][8][9]. Although some amino acid residues do demonstrate a preference for the α -helix structure, it is only marginal. For example, the most helix-preferring amino acid Glu occurs in α -helices only 59% more frequently than would occur randomly. Even Gly and Pro residues, which are not stereo chemically compatible with the α -helical conformation, are found in α -helices about 40% as often as occur randomly [10].

The α -helix is characterized by consecutive main-chain (*i, i-4*) hydrogen bonds between an amide hydrogen and a carbonyl oxygen. This pattern, however, is interrupted at the N- and C-termini, because, upon termination, no turn of the helix follows to provide additional hydrogen bond patterns. Such end effects substantially influence amino acid preference [6][11][12][13].

The position-specific preferences of particular amino acid residues are found especially at the terminal regions of the α -helices. For example, the acidic Asp and Glu residues predominate at the N-terminus, and the basic Lys, Arg, and His residues at the C-terminus, as a result of the favorable interactions of their charges with the helix dipole. Although Pro residues rarely occur in the interior of α -helices, where their unusual backbones interrupt the α -helix and cause it to kink, they frequently occur at the first N-terminal turn of the α -helix, where their particular geometry fits well. Asn, Asp, Ser, and Thr residues often occur in the first turn, where their side chains tend to form a hydrogen bond with the backbone of the third residue further along. In contrast, Gly residues occur at the carboxyl end of about a third of all α -helices, where the more flexible backbone of this residue tends to disrupt the α -helix by tending toward the 3_{10} type conformation [10].

For the interior of the α -helices, the amino acid preference is provided by the normalized frequency (the fraction of an amino acid residue occurring in the α -helices divided by its fraction in all of the proteins). This property indicates the propensity of each amino acid for forming an α -helix. Since the α -helix has a very regular structure in the interior, it is usual not to pay much attention to position specificity of the amino acid preference.

It is well known, however, that the amino acid occurrences vary with the locations in protein folding, the geometries, and the lengths of the α -helices, even in the interior region [6][8][9][14]. As a matter of fact, many α -helices are amphipathic, whereby they have predominantly non-polar residues along one side of the α -helical cylinder and polar residues along the other [15]. The two sides are usually referred to as being hydrophobic and hydrophilic, respectively. However, this definition is not adequate to analyze the position-specific amino acid preferences in the interior of the α -helices, because the two sides are defined with respect to the richness in hydrophobic and hydrophilic residues, respectively. (The statement that the hydrophobic residues preferably occur on the hydrophobic side of the α -helix makes no sense). Alternatively, the position-specific amino acid propensities in the interior of the α -helices were obtained by considering the solvent inaccessibility of those residues [6][14].

In this paper we are interested in analyzing the environment-dependent and position-specific frequencies of amino acid occurrence in α -helices, but our approach to such a problem is quite different from those described above. We chose to define local structural motifs of proteins by the Delaunay tessellation method proposed by Wako and Yamato [16]. In that method, the three-dimensional structure of a protein molecule is divided up into non-overlapping Delaunay tetrahedrons, on each vertex of which one of the comprising residues is located. The Delaunay tessellation can be performed uniquely for a given protein. To then characterize a local structure constructed by the residues on the vertices of several tetrahedrons spatially neighboring each other, a code (Delaunay code) is assigned to each tetrahedron according to the rules proposed by Wako

and Yamato, which have been slightly modified from the original ones in this paper.

We focus our attention on the interior of the α -helix, because we want to demonstrate the ability of the Delaunay code to differentiate the interior positions of the α -helix in spite of its regularity. Although the N- and C-terminal regions of the α -helix are also interesting, we will discuss them only briefly, because much research has been carried out on the terminal regions. It should be also emphasized here that the Delaunay code was devised to analyze not only the α -helix, but also various local motifs. Although we confined ourselves to analyzing the only those Delaunay codes related to the α -helix in this paper, the same approach is applicable to other codes.

As described below, the Delaunay code assigned to the tetrahedron located in the interior of an α -helix is given as FHABCDEG: FHABCDEG: x: y: FHABCDEG, where 36 kinds of codes are possible for x: y. The differences in the codes for the α -helices arising from x: y reflect this situation, whether or not some residues are located around the relevant α -helices. In other words, the environment of the α -helix can be differentiated by these 36 kinds of codes. Accordingly, we examined the frequency of amino acid occurrence on a given position of a given code (*i.e.*, a given vertex of the Delaunay tetrahedron in a given environment) in the interior of the α -helix. It should be noted that our interest is the analysis of the amino acid frequencies of occurrences with respect to its environment rather than their preferences for the formation of the α -helix instead of any other conformational states, such as an extended structure, a turn, or a random coil.

2. Methods

2.1 Local structure code

Here, we review the Delaunay tessellation and code assignment to the tetrahedrons briefly, at first. In this paper the code assignment rules are slightly changed from those defined by Wako and Yamato [16].

The three-dimensional structure of a protein molecule is represented as a set of $C\alpha$ atoms. In the previous paper, if a protein has more than one chain, each one has to be treated independently. In this study, however, we changed the code assignment rules so that two or more chains could be treated together. This modification makes it possible to assign a code to the tetrahedron consisting of residues in the interfacing regions of the two subunits.

By the Delaunay tessellation, the interior space of the protein is divided up into non-overlapping Delaunay tetrahedrons whose vertices are the $C\alpha$ atoms. Some edges of the tetrahedrons are virtual bonds connecting adjacent $C\alpha$ atoms along the polypeptide chain, and others connect two non-adjacent $C\alpha$ atoms near each other in space.

Consider a Delaunay tetrahedron T_0 . The amino acid residue number at the four vertices of T_0 are denoted as $v_1(T_0)$, $v_2(T_0)$, $v_3(T_0)$, and $v_4(T_0)$. Here we can require the suffixes to satisfy $v_1(T_0) < v_2(T_0) < v_3(T_0) < v_4(T_0)$ without losing generality.

We also consider the tetrahedrons neighboring T_0 , which share one of the facets (triangular faces) of T_0 . At most four tetrahedrons, T_5 , T_6 , T_7 , and T_8 , can possibly exist, although they do not always do so. If any such do exist, the four tetrahedrons, T_5 , T_6 , T_7 , and T_8 , are defined as sets of vertex residues, $\{v_2, v_3, v_4, v_5\}$, $\{v_1, v_3, v_4, v_6\}$, $\{v_1, v_2, v_4, v_7\}$, and $\{v_1, v_2, v_3, v_8\}$, respectively.

Furthermore, we take into account the tetrahedrons neighboring T_5 to T_8 . At most 12 more tetrahedrons can possibly exist. They are defined as $T_9 = \{v_3, v_4, v_5, v_9\}$, $T_{10} = \{v_2, v_4, v_5, v_{10}\}$, $T_{11} = \{v_2, v_3, v_5, v_{11}\}$, $T_{12} = \{v_3, v_4, v_6, v_{12}\}$, $T_{13} = \{v_1, v_4, v_6, v_{13}\}$, $T_{14} = \{v_1, v_3, v_6, v_{14}\}$, $T_{15} = \{v_2, v_4, v_7, v_{15}\}$, $T_{16} = \{v_1, v_4, v_7, v_{16}\}$, $T_{17} = \{v_1, v_2, v_7, v_{17}\}$, $T_{18} = \{v_2, v_3, v_8, v_{18}\}$, $T_{19} = \{v_1, v_3, v_8, v_{19}\}$, and

$T_{20}=\{v_1, v_2, v_8, v_{20}\}$. In this way, we can assign v_1 to v_{20} to some residues in the given protein uniquely for each tetrahedron.

We consider local structures consisting of these 20 residues. In actual fact, however, most local structures consist of less than 20 residues, because (1) some tetrahedrons do not exist, and (2) some vertices coincide with other vertices.

Then, we assign the two kinds of codes, called ST and NNT codes in the previous paper, to each tetrahedron.

The ST code is defined by arranging the vertex residue numbers v_1 to v_8 in increasing order. The ST code is a string of the suffices of v 's in that order, and is assigned to the tetrahedron T_0 . For example, if $v_8 < v_7 < v_1 < v_2 < v_5 < v_6 < v_3 < v_4$, then the ST code of the tetrahedron T_0 is 87125634. In this paper, however, we modified this code assignment rule with respect to the four points.

The first point of the modifications is to convert the figures 1 to 8 to letters of the alphabets A to H, respectively. Both the upper and lower cases are allowed to be used under the rules described below.

The second point of the modification is to distinguish the residues in separate regions along the polypeptide chain(s). For the above example, assume that the eight residues are localized into three different regions, (v_8, v_7, v_1, v_2) , (v_5, v_6) , and (v_3, v_4) , in the chain. We can also assume another case consisting of two different regions such as $(v_8, v_7, v_1, v_2, v_5)$ and (v_6, v_3, v_4) . Here, the two residues are regarded as being separated, if more than 3 residues exist between them along the same chain, or if they are located in different chains. In the examples, the differences between v_5 and v_2 and between v_3 and v_6 in the first case, and between v_5 and v_6 in the second case, should be greater than 3. The two cases cannot be distinguished by the original code assignment rules. By making use of alphabet letters, we can use the lower and upper cases alphabets alternately to distinguish the neighboring regions; thus 87125634 is, for example, converted into HGABefCD and HGABefcd for the above first and second cases, respectively.

In addition to the above examples, we introduce two more examples to explain another new rule introduced into the code assignment. One example is $v_5 < v_6 < v_8 < v_7 < v_1 < v_2 < v_3 < v_4$ comprising three regions, (v_5, v_6) , (v_8, v_7, v_1, v_2) , and (v_3, v_4) . Another is $v_8 < v_7 < v_1 < v_2 < v_3 < v_4 < v_5 < v_6$ comprising three regions, (v_8, v_7, v_1, v_2) , (v_3, v_4) , and (v_5, v_6) . The codes EFhgabCD and HGABcdEF, are assigned to these tetrahedrons, respectively, according to the rules described so far. Although in the previous paper we regarded these three codes HGABefCD, EFhgabCD, and HGABcdEF as different local structures, we intended to regard them as the same local structure in this paper. For this purpose, we have added a new rule: if there are more than one region in a code, they are arranged in the order of their sizes (the numbers of residues contained in the regions); if the sizes are equal to each other, however, they are arranged in the alphabetical order. In the examples, HGABcdEF represents all of the three codes. Owing to this modification we are able to not only unify several codes into one code, but also assign the code to a tetrahedron lying across the two chains (such an assignment was impossible according to the original rules, because a comparison of residue numbers in different chains made no sense).

The fourth point is that if the residue is missing at any of the vertices v_5 to v_8 , the lower case letter x is used to represent such vacant vertices. In the previous paper, such vertices are not included in the code. As a result any ST code consists of eight alphabets, while four to eight figures in the previous paper. This rule is introduced solely for computational convenience.

After the ST codes are assigned to all of the tetrahedrons in the protein, then they are re-assigned NNT codes taking into account the surrounding tetrahedrons. That is, the NNT code for T_0 is defined as $c(T_0): c(T_5): c(T_6): c(T_7): c(T_8)$, where $c(T_i)$ is the ST code for the tetrahedron T_i . This procedure is the same as the previous one. Hereinafter, the NNT code is simply referred to as the Delaunay code.

2.2 Codes for α -helix

One of the typical codes for the interior of the α -helix is FHABCDEG: FHABCDEG: c(T₆): c(T₇): FHABCDEG (which corresponds to the code 68123457: 68123457: c(T₆): c(T₇): 68123457 in the previous paper). The correspondence between residues and vertex positions are given in Table 1 and Figure 1. This type of code is the most abundant in the proteins. There are 36 possible codes for c(T₆): c(T₇) as shown in Table 2. Both c(T₆) and c(T₇) have the string ABEHCD, followed by F/f, G/g, and/or x in various combinations.

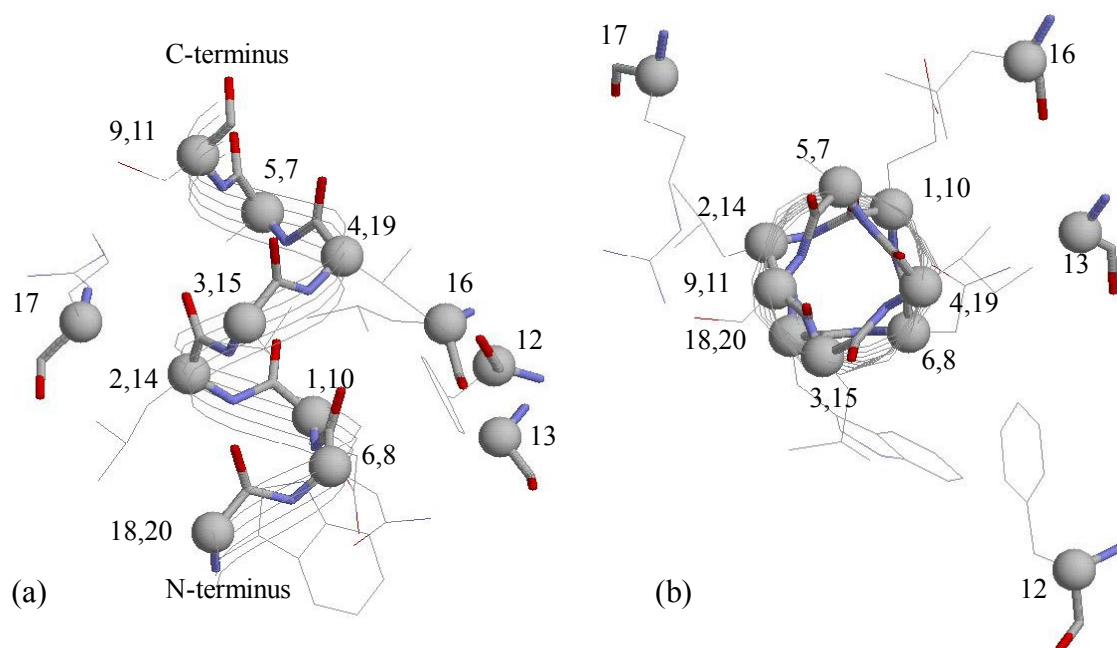


Figure 1. Vertex numbers in the interior of an α -helix with the code FHABCDEG: FHABCDEG: c(T₆): c(T₇): FHABCDEG.

α atoms are represented by the balls. Whether or not the vertex residues 12, 13, 16, and 17 exist is reflected in c(T₆) and c(T₇) (see Table 1). Two views are shown: (a) parallel and (b) perpendicular to the helical axis.

Table 1. Codes for the interior of α -helices, and the correspondence between residues and vertex positions (FHABCDEG: FHABCDEG: c(T₆): c(T₇): FHABCDEG)

Residue	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	others*			
Vertex	V ₁₈ V ₂₀	V ₆ V ₈	V ₁ V ₁₀	V ₂ V ₁₄	V ₃ V ₁₅	V ₄ V ₁₉	V ₅ V ₇	V ₉ V ₁₁	V ₁₂	V ₁₃	V ₁₆	V ₁₇
c(T ₀)		F,H	A	B	C	D	E,G					
c(T ₅)			F,H	A	B	C	D	E,G				
c(T ₆)		A	B	E,H	C	D			F/f,x	G/g,x		
c(T ₇)			A	B	E,H	C	D				F/f,x	G/g,x
c(T ₈)	F,H	A	B	C	D	E,G						

* The residues separated by more than 3 residues from the residues i-2 to i+5 along the chain are classified as others.

Table 2. Possible codes of $c(T_6)$ and $c(T_7)$ for the interior of α -helix.

$c(T_6)$	$c(T_7)$	No. of structures observed		Surrounding residues [*]
		intra-chain	contact surface	
ABEHCDfG	ABEHCDfG	2346	414	1111
ABEHCDfg	ABEHCDfg	1425	157	1111
ABEHCDfx	ABEHCDfx	1987	160	1110
ABEHCDgf	ABEHCDgf	370	39	1111
ABEHCDgx	ABEHCDgx	381	82	1101
ABEHCDxx	ABEHCDxx	1000	60	1100
ABEHCDfG	ABEHCDfG	1175	159	1111
ABEHCDfg	ABEHCDfg	732	93	1111
ABEHCDfx	ABEHCDfx	869	44	1110
ABEHCDgf	ABEHCDgf	189	23	1111
ABEHCDgx	ABEHCDgx	582	78	1101
ABEHCDxx	ABEHCDxx	1292	40	1100
ABEHCDfG	ABEHCDfG	503	94	1011
ABEHCDfg	ABEHCDfg	562	76	1011
ABEHCDfx	ABEHCDfx	376	33	1010
ABEHCDgf	ABEHCDgf	163	21	1011
ABEHCDgx	ABEHCDgx	1203	108	1001
ABEHCDxx	ABEHCDxx	2208	66	1000
ABEHCDfG	ABEHCDfG	414	41	1111
ABEHCDfg	ABEHCDfg	157	15	1111
ABEHCDfx	ABEHCDfx	320	14	1110
ABEHCDgf	ABEHCDgf	74	8	1111
ABEHCDgx	ABEHCDgx	127	16	1101
ABEHCDxx	ABEHCDxx	273	14	1100
ABEHCDfG	ABEHCDfG	1895	122	0111
ABEHCDfg	ABEHCDfg	786	46	0111
ABEHCDfx	ABEHCDfx	802	54	0110
ABEHCDgf	ABEHCDgf	253	14	0111
ABEHCDgx	ABEHCDgx	211	14	0101
ABEHCDxx	ABEHCDxx	226	13	0100
ABEHCDfG	ABEHCDfG	1034	73	0011
ABEHCDfg	ABEHCDfg	985	37	0011
ABEHCDfx	ABEHCDfx	315	19	0010
ABEHCDgf	ABEHCDgf	253	16	0011
ABEHCDgx	ABEHCDgx	2185	70	0001
ABEHCDxx	ABEHCDxx	1889	0	0000

*The existence and absence of residues on the vertices v_{12} , v_{13} , v_{16} , and v_{17} are indicated by 1 and 0, respectively. For example, 1001 means that residues exist on v_{12} and v_{17} , but not on v_{13} and v_{16} .

2.3 Statistical analysis of amino acid occurrence

Let $f_{j,c,b}$ be the frequency of occurrence of amino acid j ($j = 1, 2, \dots, 20$) at a given site b (a given vertex of the tetrahedron) of a given code c , and define $\mathbf{f}_{c,b} = (f_{1,c,b}, \dots, f_{20,c,b})$. This is a key property in this study. For the interior of the α -helix, $c = 1, 2, \dots, 36$. As for the vertices, we

confine ourselves to the four vertices v_1 to v_4 in the central tetrahedron T_0 in the following statistics (i.e., $b = 1, 2, 3, 4$).

Accordingly, there are $36 \times 4 = 144$ points for $\mathbf{f}_{c,b}$ in the 20-dimensional space. Their distribution can be characterized by the principal component analysis as follows.

Mean $\langle f_j \rangle$ and standard deviation s_j for a given amino acid j are calculated over the 4 vertices of the $m=36$ codes. The normalized frequency $x_{j,c,b}$ is then introduced as

$$x_{j,c,b} = \frac{f_{j,c,b} - \langle f_j \rangle}{s_j}. \quad (1)$$

For the sake of convenience, these values are gathered up in a matrix:

$$\mathbf{X} = \begin{bmatrix} x_{1,1,1} & \cdots & x_{20,1,1} \\ x_{1,1,2} & \cdots & x_{20,1,2} \\ \vdots & \cdots & \vdots \\ x_{1,m,4} & \cdots & x_{20,m,4} \end{bmatrix}. \quad (2)$$

A correlation between amino acid i and j , $C_{ij} = \frac{1}{4m} \sum_{b=1}^4 \sum_{c=1}^m x_{i,c,b} x_{j,c,b}$, is expressed in a matrix form:

$$\mathbf{C} = \frac{1}{4m} \mathbf{X}' \mathbf{X}. \quad (3)$$

where \mathbf{X}' is a transpose matrix of \mathbf{X} . To make a principal component analysis, the eigenvalues and eigenvectors of matrix \mathbf{C} are calculated:

$$\mathbf{C} \mathbf{u}_k = \lambda_k \mathbf{u}_k, \quad (4)$$

where eigenvector \mathbf{u}_k satisfies the orthonormal condition such that $\mathbf{u}_i \mathbf{u}_j = 1$ if $i = j$, and $\mathbf{u}_i \mathbf{u}_j = 0$ if $i \neq j$. We assume $\lambda_1 > \lambda_2 > \cdots > \lambda_{20}$. Then \mathbf{u}_k is a unit vector along the k th principal component axis. Since the data projected on the k th axis have the standard deviation $\sigma_{kj} = \sqrt{\lambda_k} u_{kj} s_j$, the following property is convenient to see the frequency of amino acid occurrence along the k th axis:

$$f_j^* = \langle f_j \rangle \pm \sigma_{kj}. \quad (5)$$

where u_{kj} is the j th component of \mathbf{u}_k .

3. Results

3.1 Structure data set

In this study, we used the structure data set of 682 representative protein chains having less than 25% homology with each other selected from Protein Data Bank [17][18][19]. Membrane proteins are not included in the data set. If an entry in the PDB contains two or more chains, the Delaunay tessellation and code assignment were carried out for the system including all the chains in the entry, even if only one of the chains is included in the representative structure data set. The tetrahedrons are classified into three categories: intra-chain (vertices v_1 to v_8 are in a representative chain),

contact surface (v_1 to v_4 are in a representative chain, but at least one of v_5 to v_8 is in another chain), and inter-chain (some of v_1 to v_4 are in a representative chain, but at least one of them is in another chain).

The numbers of tetrahedrons categorized into intra-chain and contact surface among the 682 representative protein chains for the codes related to the interior of the α -helix are shown in Table 2. No tetrahedron is observed in the inter-chain category for these codes as a matter of course. Only the intra-chain tetrahedrons were used in the statistical analyses described below. The contact-surface tetrahedrons were analyzed separately. The results are also given below briefly, since the number of tetrahedrons in this category is too small to make a reliable statistical analysis.

3.2 Principal component analysis

At first we examine the position-independent properties. The mean frequency of amino acid occurrence and its standard deviation over vertices v_1 to v_4 of the 36 codes related to the interior of the α -helix are given in Figure 2(a). Ala (12.8%) and Leu (12.2%) are the most abundant, followed by Glu, Val, Lys, Ile, and Arg (6 to 8 %). In contrast, the percentages for His, Trp, Cys, and Pro are less than 2 %. These results are generally well correlated with other analyses of amino acid preferences for formation of the α -helix [1][4], although the preference is usually defined as the fraction of the residues of each amino acid that occur in an α -helix, divided by the fraction of its random occurrence.

The standard deviations for Leu, Glu, Ala, and Lys (5.2, 4.6, 4.0, and 3.7 %, respectively) are larger than 3.0 %. The fact that the standard deviation of Leu is much larger than that of Ala, in spite of their nearly equal mean values, indicates that the occurrence of Leu is more biased (*i.e.*, more dependent on the environment of vertices) than is Ala (*i.e.*, more independent). In the same context, the occurrence of Glu and Lys is more biased than that of Val, Ile, and Arg.

The top four largest eigenvalues, λ_1 to λ_4 , are 11.1, 2.7, 1.3, and 0.9, respectively. Their contribution to variance, *i.e.*, $\lambda_k / 20$, are 55.6 %, 13.5 %, 6.7 %, and 4.3 %, respectively. The top two eigenvalues λ_1 and λ_2 contribute as much as 70% to variance.

f_j^* defined by Equation 5 is plotted for the first and second principal components in Figures 2b and 2c, respectively. The large negative values of σ_{kj} for Glu (-4.2 %) and Lys (-3.3 %), and the large positive one for Leu (4.7 %) in the first principal component (the line with solid square symbols in Figure 2b; although the negative and positive signs of σ_{kj} are interchangeable, *i.e.*, the two lines with the solid square and open triangle symbols in Figure 2b are interchangeable, we refer to the signs of σ_{kj} in this manner for convenience sake) indicate that Glu, Lys, and Leu are characteristic amino acids for differentiating the two type of positions in the interior of the α -helix, *i.e.*, the hydrophilic ($\sigma_{kj} < 0$) and hydrophobic ($\sigma_{kj} > 0$) ones. Asp (-2.5 %), Arg (-1.9 %), and Gln (-1.8 %) also have relatively larger negative σ_{kj} values, and Ile (2.6 %), Val (2.5 %), Ala (1.9 %), and Phe (1.6 %) have relatively larger positive σ_{kj} values.

In the second principal component, large negative values of σ_{kj} are observed for Leu (-1.7 %), Lys (-1.1 %), and Arg (-0.8 %), and large positive ones for Ala (3.0 %), Gly (1.3 %), and Ser (0.7 %) (the line with solid square symbols in Figure 2c; the negative and positive signs of σ_{kj} are also interchangeable as described above). The second principal component indicates another kind

of characteristics to differentiate the positions in the interior of the α -helix, *i.e.*, the preference for the amino acids with larger ($\sigma_{kj} < 0$) and smaller ($\sigma_{kj} > 0$) sidechains. In actual fact, those values for the amino acids with aromatic rings, His, Tyr, Phe, and Trp, are negative, and those for Pro, Val, Thr and Cys are positive, although their absolute values are much smaller than the above residues.

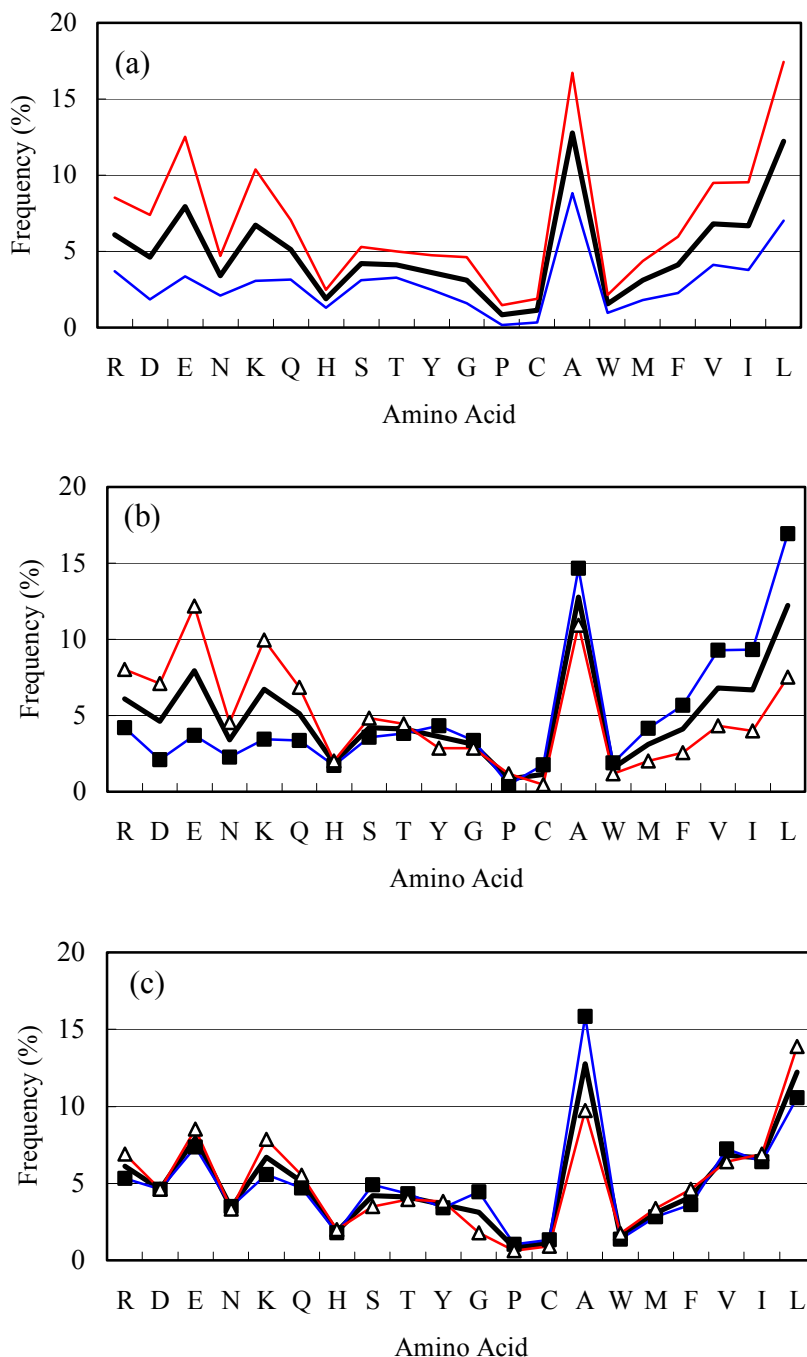


Figure 2. Frequency of amino acid occurrence in the interior of the α -helix.

(a) $\langle f_j \rangle$ (black line) and $\langle f_j \rangle \pm s_j$ (colored lines); (b) and (c) f_j^* (Equation 5) for the first and second principal axes, respectively, (colored lines) and $\langle f_j \rangle$ (black line).

surrounding residues exist. At the vertices v_9 and v_{18} , both hydrophobic and hydrophilic amino acids occur with nearly equal frequencies. These vertices are relatively independent from the surrounding residues, and are located on the opposite sides of the α -helix cylinder against v_1 and v_4 (Figure 1). It frequently occurs that when hydrophobic amino acids are preferable on one side, hydrophilic amino acids are preferable on the other side. This fact may be reflected in the statistics for these vertices. This also holds in Figures 4b and 4c.

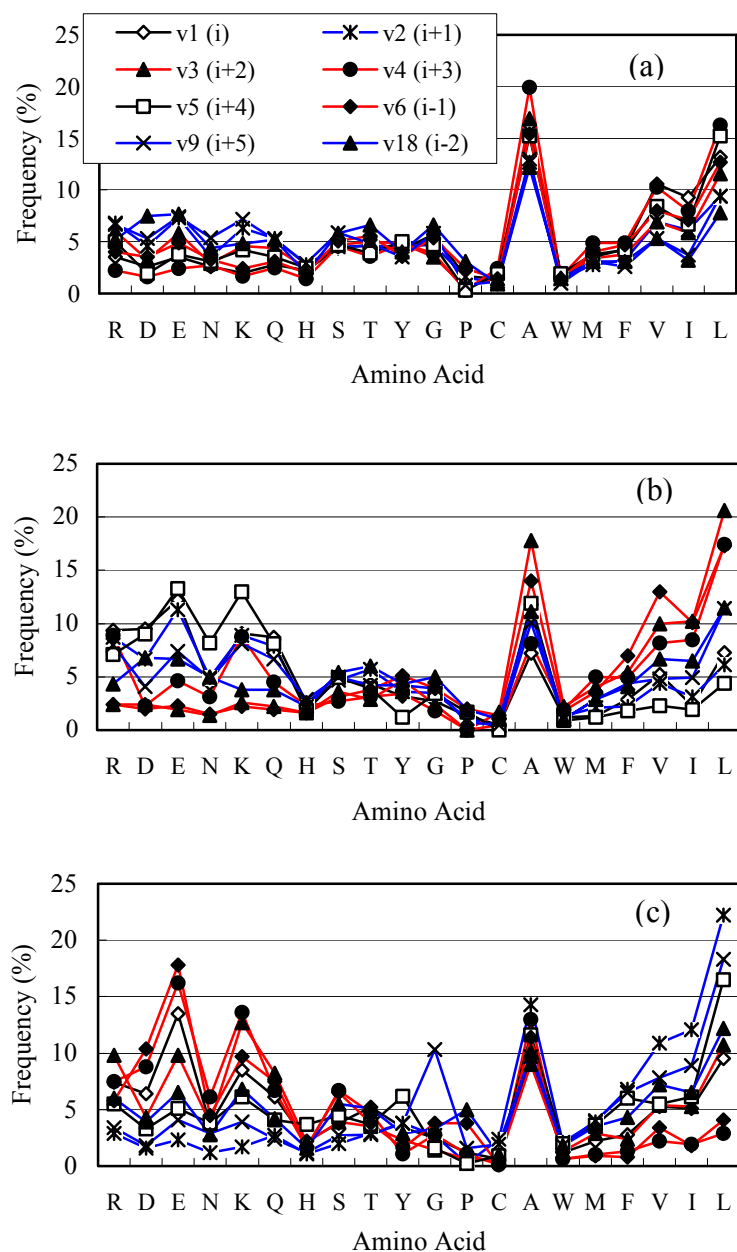


Figure 4. Frequencies of amino acid occurrences $f_{c,b}$ at eight vertices corresponding to residues $i-2$ to $i+5$ in the interior of the α -helix.

The data are shown for the three kinds of codes: (a) $c(T_6)$: $c(T_7)$ =ABEHCDfG: ABEHCDfG, (b) ABEHCDfG: ABEHCDxx, and (c) ABEHCDxx: ABEHCDgx.

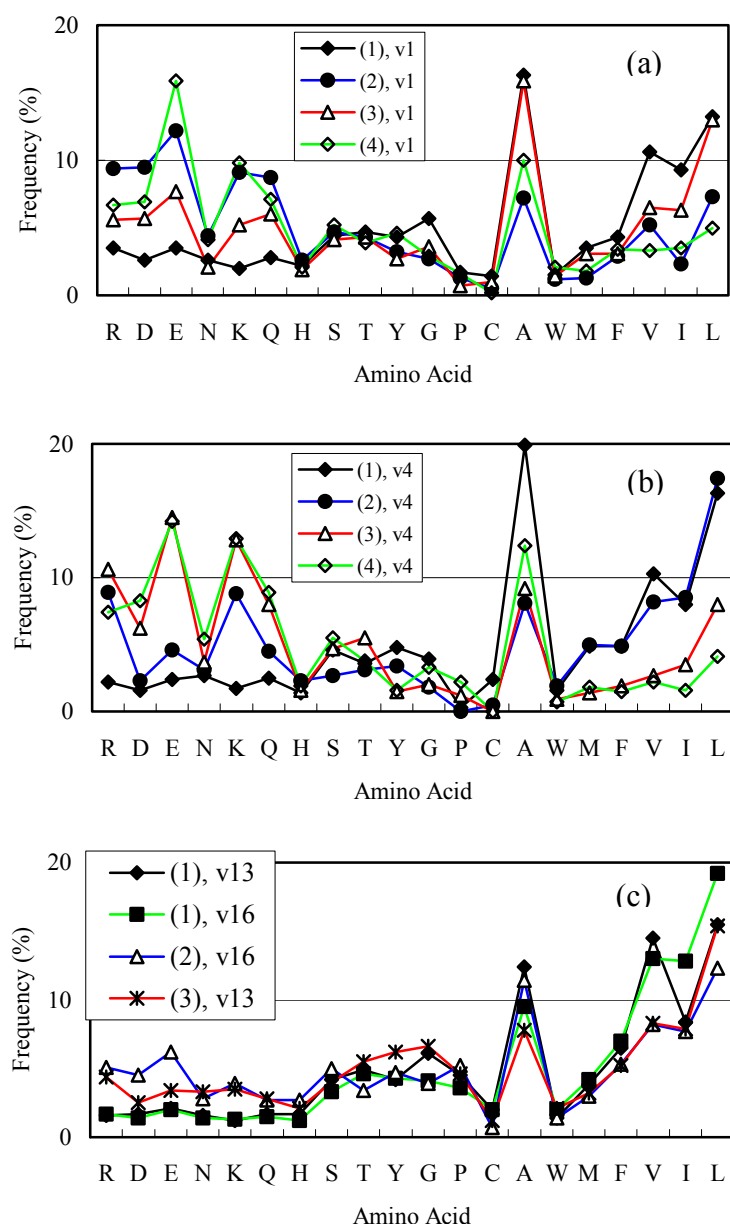


Figure 5. Frequencies of amino acid occurrences $f_{c,b}$ at the vertices, v_1 , v_4 , v_{13} , and v_{16} in the interior of the α -helix.

The data are shown for the four kinds of codes: (1) $c(T_6): c(T_7) = ABEHCDfG: ABEHCDfG$, (2) $ABEHCDfG: ABEHCDxx$, (3) $ABEHCDxx: ABEHCDfg$, and (4) $ABEHCDxx: ABEHCDxx$.

$f_{c,b}$ for v_1 and v_4 of these four codes are given in (a) and (b), respectively. $f_{c,b}$ for v_{13} and v_{16} of the three codes, (1) to (3), are given in (c).

To interpret the results shown in Figures 4b and 4c, we have to take into consideration the vertices composing the tetrahedrons T_{12} , T_{13} , T_{16} , and T_{17} ; *i.e.*, $T_{12} = \{v_3, v_4, v_6, v_{12}\}$, $T_{13} = \{v_1, v_4, v_6, v_{13}\}$, $T_{16} = \{v_1, v_4, v_7, v_{16}\}$, and $T_{17} = \{v_1, v_2, v_7, v_{17}\}$ (see also Figure 1). In Figure 4b, v_3 , v_4 , and v_6 prefer hydrophobic amino acids owing to the existence of v_{12} and v_{13} , whereas v_1 , v_2 , and v_5 prefer

hydrophilic ones owing to the absence of v_{16} and v_{17} . Conversely, in Figure 4c, v_2 and v_5 ($=v_7$) prefer hydrophobic amino acids owing to the existence of v_{17} , whereas v_4 and v_6 prefer hydrophilic ones owing to the absence of v_{12} , v_{13} , and v_{16} . v_1 is ambivalent.

Gly at v_9 in Figure 4c shows remarkable high frequency. Since Gly is frequently found in the C-cap region of the α -helix, it may indicate that the situation represented by this code appears more often near the C-terminal of the α -helix.

To illustrate that the frequencies $f_{c,b}$ differ depending on the codes (i.e., environment) even for the same vertex position, some examples are given in Figures 5a and 5b. For the four codes, (1) $c(T_6): c(T_7) = \text{ABEHCDfG}: \text{ABEHCDfG}$, (2) $\text{ABEHCDfG}: \text{ABEHCDxx}$, (3) $\text{ABEHCDxx}: \text{ABEHCDfg}$, and (4) $\text{ABEHCDxx}: \text{ABEHCDxx.}$, the frequencies $f_{c,b}$ on the vertices v_1 and v_4 are shown. As for the four residues surrounding the α -helix, v_{12} , v_{13} , v_{16} , and v_{17} , all of them exist in (1), only v_{12} and v_{13} in (2), only v_{16} and v_{17} in (3), and none of them in (4). v_1 is affected by the existence of the pair of v_{16} and v_{17} , and v_4 by the pair of v_{12} and v_{13} . As a result, v_1 of (1) and (3) prefers hydrophobic amino acids (Figure 5a), and so does v_4 of (1) and (2) (Figure 5b). Both v_1 and v_4 of (4) prefer hydrophilic amino acids.

In Figure 5c, the frequencies $f_{c,b}$ for the surrounding vertices v_{13} and v_{16} are shown. In any case the hydrophobic amino acids are strongly preferable, because they interact with the residues in the α -helix.

Through Figures 4 and 5, the Ala residue is remarkable. While Ala behaves essentially like the hydrophobic amino acids, it frequently appears at vertices preferring hydrophilic amino acids. The small sidechain and strong preference for forming the α -helix are considered to make it possible.

The results shown so far can be examined from the viewpoints of principal components. The first and second principal components for the frequencies $f_{c,b}$, i.e., the projections on the first and second principal axes, were calculated for each of the 36 codes. Only the data for the vertices (a) v_1 and (b) v_4 are plotted in Figure 6. To clarify the results, the codes were divided into six and five groups for v_1 and v_4 , respectively, taking into account the existence of residues at the four vertices, v_{12} , v_{13} , v_{16} , and v_{17} (see the legend for Figure 6). In the first principal axis (horizontal one), the large positive (negative) value indicates the preference for the hydrophobic (hydrophilic) amino acids. In the second principal axis (vertical one), the large positive (negative) values indicate the preference for amino acids with smaller (larger) sidechains.

Since $T_{13}=\{v_1, v_4, v_6, v_{13}\}$ and $T_{16}=\{v_1, v_4, v_7, v_{16}\}$, both v_1 and v_4 are affected by v_{13} and v_{16} . On the other hand, $T_{12}=\{v_3, v_4, v_6, v_{12}\}$ and $T_{17}=\{v_1, v_2, v_7, v_{17}\}$ indicate that v_{12} can affect v_4 but not v_1 , whereas v_{17} can affect v_1 but not v_4 . Generally speaking, Figure 6 shows that in the presence of such influential residues, the hydrophobic amino acids with smaller sidechains (open symbols in Figure 6) are preferable, and that in the absence of them the hydrophilic amino acids with larger sidechains (closed symbols) are preferable. However, in some cases (Group 5 for v_1 and Group 4 for v_4 , for which the symbol * is used in both (a) and (b) of Figure 6), where some of the influential residues exist but others do not, hydrophobic amino acids with larger sidechains are preferable.

3.4 Tetrahedron on the contact surface

The tetrahedrons classified as contact surfaces consist of the vertices v_1 to v_4 in the same chain and at least one of the vertices v_5 to v_8 in a different chain. Although such tetrahedrons assigned to the α -helix-related codes are found in the protein structure set considered here, the number of such tetrahedrons is too small to obtain reliable statistical analysis results (see Table 2). For the cases of $c(T_6): c(T_7)= \text{ABEHCDfG}: \text{ABEHCDfG}$ and $\text{ABEHCDfG}: \text{ABEHCDfg}$, where the numbers of the

data are relatively larger than the others, the differences of $f_{c,b}$ between tetrahedrons classified as intra-chain and contact surface are plotted for some vertices in Figure 7. In Figure 7a, the differences are plotted for the vertices v_1 , v_4 , and v_6 facing surrounding residues v_{12} , v_{13} , v_{16} , and v_{17} , some of which are in a different chain. In contrast, in Figure 7b, the differences are plotted for the vertices v_2 and v_9 facing the interior residues in their own chain. The frequencies of the hydrophilic amino acids increase in Figure 7a and so do those of the hydrophobic amino acids in Figure 7b, although such changes are too subtle to be asserted.

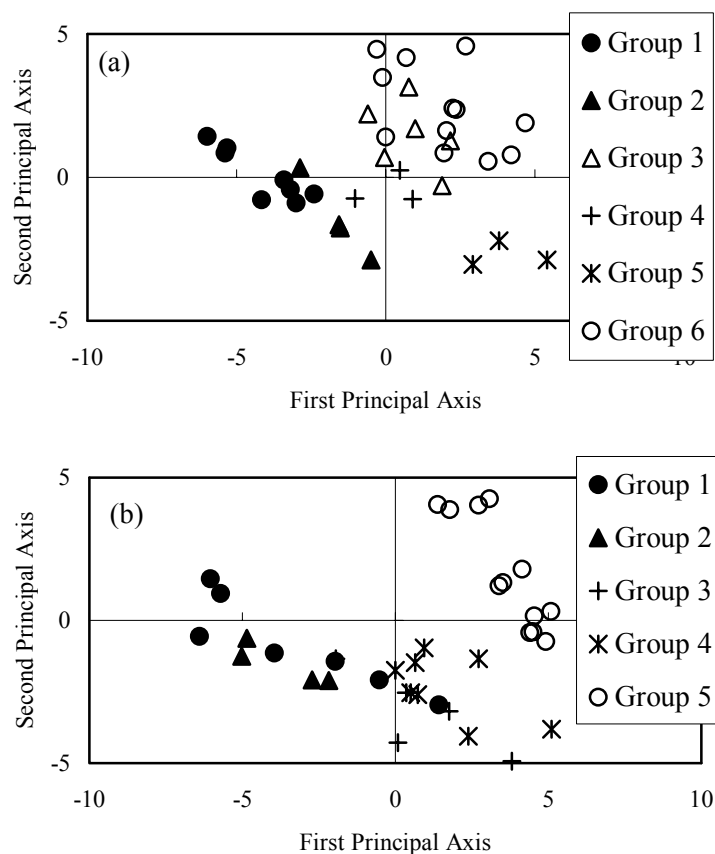


Figure 6. The first and second principal components for the frequencies $f_{c,b}$ at vertices v_1 (a) and v_4 (b).

Grouping of the 36 codes given in Table 2 is as follows: (a) Group 1 (0000, 0001, 0100, 1000, 1001, 1100), Group 2 (1010, 1110), Group 3 (0011, 1011), Group 4 (1101), Group 5 (0010, 0101, 0110) and Group 6 (0111, 1111); (b) Group 1 (000, 0001, 1001, 1000, 1100), Group 2 (0011, 0010), Group 3 (1101, 0101, 0100), Group 4 (0111, 0110, 1011, 1010), and Group 5 (1110, 1111) (see the column “Surrounding residues” in Table 2 for the key to the four-figure notation).

3.5 Cap regions

The cap regions are very interesting in the analysis of the α -helix. In this paper, however, we analyzed only the tetrahedrons characterized by the codes $c(T_1) = c(T_5) = \text{FHABCDEG}$ but $c(T_8) \neq \text{FHABCDEG}$, and $c(T_1) = c(T_8) = \text{FHABCDEG}$ but $c(T_5) \neq \text{FHABCDEG}$, which correspond to the N- and C-caps, respectively, while $c(T_1) = c(T_5) = c(T_8) = \text{FHABCDEG}$ in the interior of the

α -helix. Either $c(T_8)$ or $c(T_5)$ not equal to FHABCDEG is an indication that the N- or C-terminal, respectively, of the α -helix is distorted.

In Table 3, the codes, for which more than 60 data entries are found in the structural data set used in this study, are shown. The correspondence between the residue and vertex positions is given in Table 4. In Table 4, the helix positions for the helix capping defined by Aurora and Rose [12][13] are also given. The frequencies $f_{c,b}$ at some vertices for (a) the N-cap with the code $c(T_5)$: $c(T_6)$: $c(T_7)$: $c(T_8) = \text{FHABCDEG}$: FGABEHCD : ABEHCDxx : ABCDEGxx and (b) the C-cap with the code FHABCDCxx : ABEHCDgx : ABEHCDfG : FHABCDEG are plotted in Figure 8.

Aurora and Rose pointed out the significance of the interactions of N3 and N4 (*i.e.*, v_3 and v_4) with N' or N'' (v_{12} and v_{13}) and the significance of Nc (v_6) for the N-cap. In Figure 8a, hydrophobic amino acids occur more frequently at vertices v_3 , v_4 , and v_{13} , while Asp, Glu, and Gln also occur frequently at v_3 and Lys and Gln at v_4 . The occurrence of Asp, Ser, Thr and Pro are remarkably higher at v_6 .

For the C-cap, Aurora and Rose revealed that Gly is rich at C' (v_5), and that interactions of C3 (v_1) with the residues following C' (not included in the codes related to the C-cap considered here) is important. In Figure 8b, Gly is exceptionally abundant at v_5 , and a hydrophobic residue, especially Leu, is found much more often at v_1 .

These results for the N-cap and C-cap generally agree well with the analyses by Aurora and Rose.

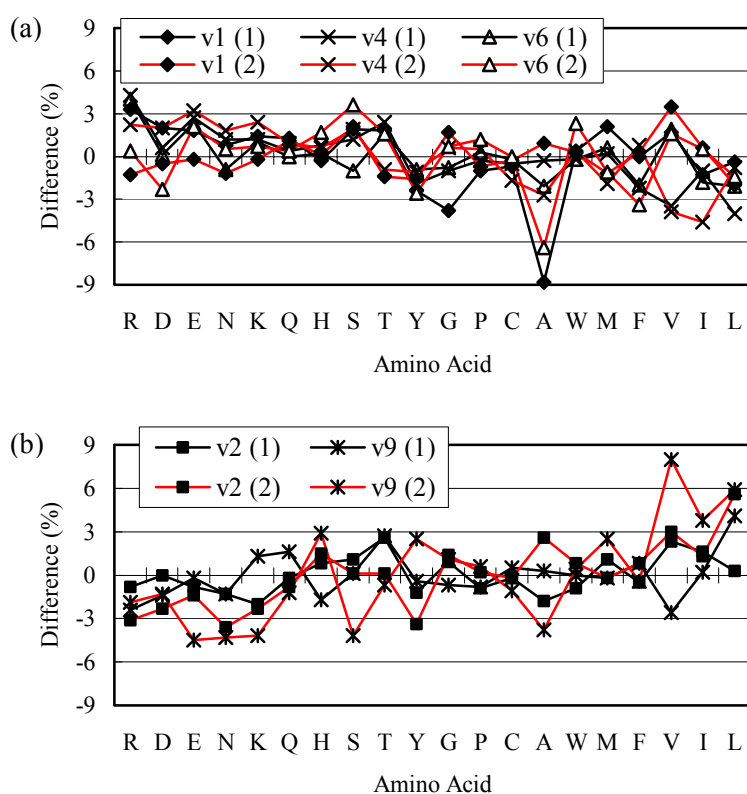


Figure 7. Differences in $f_{c,b}$ for contact-surface tetrahedrons from those for intra-chain tetrahedrons.

The results are shown for the six vertices, v_1 , v_2 , v_4 , v_6 , and v_9 , of the two codes, (1) $c(T_6)$: $c(T_7) = \text{ABEHCDfG}$: ABEHCDfG and (2) ABEHCDfG : ABEHCDfg .

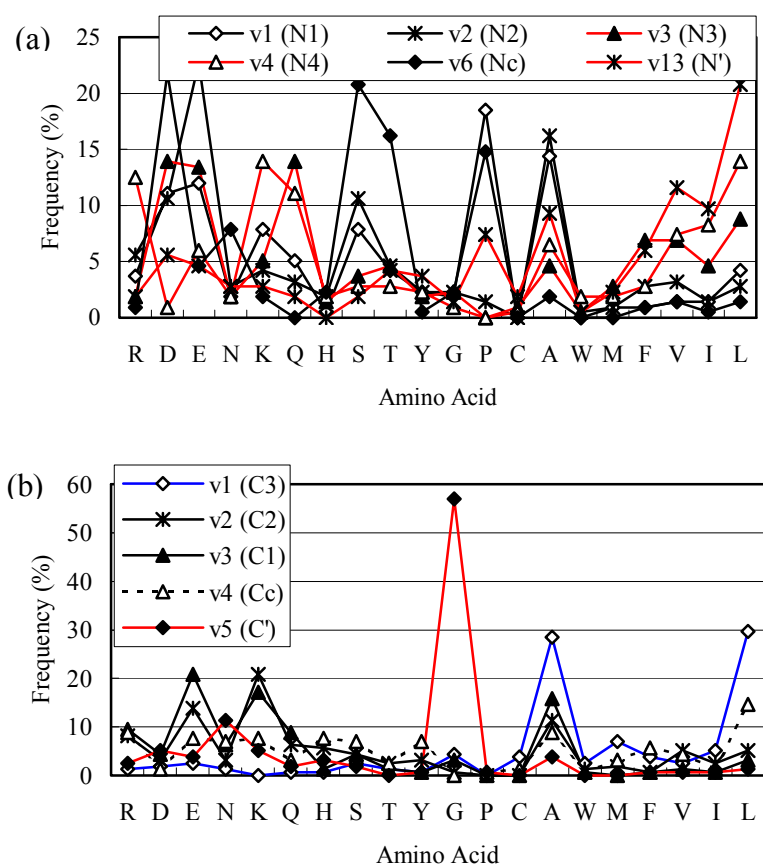


Figure 8. Amino acid frequencies $f_{c,b}$ for N-cap and C-cap.

(a) $c(T_5)$: $c(T_6)$: $c(T_7)$: $c(T_8)$ = FHABCDEG: FGABEHCD: ABEHCDxx: ABCDEGxx (N-cap) and
 (b) FHABCDxx: ABEHCDgx: ABEHCDFG: FHABCDEG (C-cap).

Table 3. Codes for N-cap and C-cap of α -helices

	$c(T_5)$	$c(T_6)$	$c(T_7)$	$c(T_8)$	No. of structures observed
N-cap	<u>FHABCDEG</u>	FABEHCDg	ABEHCDfx	ABCDEGxx	94
	<u>FHABCDEG</u>	FGABEHCD	ABEHCDfG	ABCDEGhx	92
	<u>FHABCDEG</u>	FGABEHCD	ABEHCDfG	ABCDEGxx	77
	<u>FHABCDEG</u>	FGABEHCD	ABEHCDfg	ABCDEGxx	127
	<u>FHABCDEG</u>	FGABEHCD	ABEHCDfx	ABCDEGxx	196
	<u>FHABCDEG</u>	FGABEHCD	ABEHCDxx	ABCDEGfx	124
	<u>FHABCDEG</u>	FGABEHCD	ABEHCDxx	ABCDEGxx	216
C-cap	FHABCDEg	ABEHCDxx	ABEHCDgx	<u>FHABCDEG</u>	62
	FHABCDxx	ABEHCDfG	ABEHCDFG	<u>FHABCDEG</u>	87
	FHABCDxx	ABEHCDfg	ABEHCDFG	<u>FHABCDEG</u>	82
	FHABCDxx	ABEHCDgx	ABEHCDFG	<u>FHABCDEG</u>	158
	FHABCDxx	ABEHCDxx	ABEHCDFG	<u>FHABCDEG</u>	122
	HABCDFxx	ABEHCDgx	ABHCEDGf	<u>FHABCDEG</u>	61

Table 4. Codes for N-cap and C-cap and correspondence between residue and vertex positions.

N-cap (FHABCDEG: FHABCDEG: FGABEHCD: ABEHCD??: ABCDEG??) *																					
Residue	i-4	i-3	i-1	i	i+1	i+2	i+3	i+4	i+5	others											
Vertex	v ₁₂	v ₁₃	v ₆	v ₈	v ₁	v ₁₀	v ₂	v ₁₄	v ₃	v ₁₅	v ₄	v ₁₉	v ₅	v ₇	v ₉	v ₁₁	v ₁₆	v ₁₇	v ₁₈	v ₂₀	
c(T ₀)			F,H	A	B	C	D	E,G													
c(T ₅)				F,H	A	B	C	D	E,G												
c(T ₆)	F,G		A	B	E,H	C	D								F/f,x	G/g,x					
c(T ₇)				A	B	E,H	C	D											F/f,x	H/h,x	
c(T ₈)			A	B	C	D	E,G														
Helix position	N'	N''	Nc	N1	N2	N3	N4	N5													

C-cap (FHABCDEG: FHABCDxx: ABEHCD??: ABEHCDFG: FHABCDEG) *																					
Residue	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	others	none											
Vertex	v ₁₈	v ₂₀	v ₆	v ₈	v ₁	v ₁₀	v ₂	v ₁₄	v ₃	v ₁₅	v ₄	v ₁₉	v ₅	v ₇	v ₁₆	v ₁₇	v ₁₂	v ₁₃	v ₉	v ₁₁	
c(T ₀)			F,H	A	B	C	D	E,G													
c(T ₅)				F,H	A	B	C	D											E/e,x	G/G,x	
c(T ₆)			A	B	E,H	C	D								F/f,x	G/g,x					
c(T ₇)				A	B	E,H	C	D	F,G												
c(T ₈)	F,H		A	B	C	D	E,G														
Helix position	C5	C4	C3	C2	C1	Cc	C'	C''													

* Symbol ? represents either F/f, G/g, or x.

4. Discussion

The preference for amino acids to form an α -helix is usually defined as the ratio of the frequency of a given amino acid found in the α -helical state to its frequency found in all the possible states. In this paper we examined the frequencies of occurrence of 20 amino acids for a given position of the α -helix in a given environment, which is specified with respect to the vertex position of a Delaunay tetrahedron and the Delaunay codes assigned to it according to the rules proposed by Wako and Yamato [16]. The former parameters are useful to predict secondary structural locations in amino acid sequences. In other words, they are useful, when we are interested in the conformational state that a given residue will take in a given sequence. In contrast, when we adopt the latter parameters, we can get information about which amino acids will occur preferably

at a given position of the α -helix in a given environment. Since the codes are assigned so as to reflect the environment of the relevant position, we can regard them as being the environment-dependent and position-specific amino acid frequencies.

Assume that we have a protein whose three-dimensional structure is known. We can assign the Delaunay code to every Delaunay tetrahedron after the Delaunay tessellation. If we want to substitute a different amino acid for some amino acid residue in the protein, we can examine the local structures containing tetrahedrons with the same code in other protein structures. The statistical analyses described here can provide helpful information about which amino acids can be candidates for the substitution, which is environment dependent and position specific.

The knowledge based 3D-1D compatibility method (or threading method) [20][21][22][23][24] also takes into account the influence from surrounding residues to assess the fitness of each residue in a given amino acid sequence for various 3D structures given as templates. The assessment of each residue is environment dependent and position specific. The differences between our method and the 3D-1D methods are mainly with respect to the aims for their use. While the 3D-1D method is used for a protein whose 3D structure is unknown, our method can be used only for the protein whose 3D structure is known. While the 3D-1D method intends to predict the 3D structure from a protein's amino acid sequence, we do not do so. We want to find the same local structures in various proteins as those occurring in the query protein. The point in the search of the local structures is that not only the local structure, but also its environment, can be checked by means of the Delaunay code. From the collection of the given local structures we can characterize them by analyzing the frequencies in it of amino acid occurrences and so on.

As for the environment-dependent amino acid substitution tables, they were constructed from structural alignment data of homologous proteins by Overington et al. [25][26]. In constructing the tables, the conformational states (defined as combinations of secondary structures, buried/exposed, hydrogen bond formation, and so on) are taken into account. These tables show that the substitution patterns depend on the conformational states. Wako and Blundell [27][28] used these tables for their prediction of the secondary-structure and solvent-accessibility classes. They emphasized the significance of the position-dependent (or conformational-state dependent) information on amino acid substitution patterns.

In connection with the environment-dependent amino acid substitution tables, it should be emphasized again for the method discussed in this paper that the conformational states and environment can be taken into account through the Delaunay codes. The classification by our method can reflect more precisely where the local structure is located and does not require the collection and alignment of homologous proteins. Although we confine ourselves with the α -helix in this paper, the same method can be applied to any of the local structure motifs defined by Delaunay codes. In other words the present method makes it possible to analyze the amino acid frequencies of occurrence for a structure without being restricted to conventional conformational states, such as α -helix, β -structure, turn, and so on.

One of the problems with the Delaunay code is its sensitivity. That is, the number of possible Delaunay codes is enormous. Although, for the more regularly structured α -helix, the number of the codes is relatively limited, the codes for other structures are full of variety. It is necessary to cluster together the codes for similar local structures. Another problem is that the local structure specified by a Delaunay code is relatively small. The extension to a larger local structure is possible with the Delaunay codes, but the process is very intricate, because the Delaunay code reduces the information about 3D structure into a 1D sequence of digits. Generally speaking, it is inevitable that some amount of information will be lost in the reduction. Thus, we do not expect that our method will be applicable for all-around use. Our method is especially useful for problems such as those presented here and those in the first paper [16].

Finally, we summarize the results obtained in this paper for the α -helix.

(1) The interior region of the α -helix represented by the Delaunay code FHABCDEG: FHABCDEG: x: y: FHABCDEG is examined. There are 36 codes of this type. The differences between the codes reflect the existence or absence of residues surrounding the α -helix.

(2) The major factors concerning the occurrence of the 20 amino acids in the interior of an α -helix are (a) preference for α -helix formation, (b) hydrophobicity and hydrophilicity, and (c) sizes of the amino acid sidechains.

(3) The above factors (b) and (c) are environment-dependent and position-specific. The codes used in this paper can represent the environment of the α -helix, and the statistics based on the codes can provide the position-specific frequencies of amino acid occurrence.

(4) Ala is a notable amino acid in the α -helix. Its behavior is essentially that of a hydrophobic amino acid. To some extent, however, it occurs at the vertices preferable for hydrophilic amino acids.

(5) The frequencies of amino acid occurrences on the α -helix surface in contact with the residues in a different chain are examined by analyzing tetrahedrons lying across the two chains. Although the difference in the frequencies from those in the one chain are very subtle, and the data are insufficient for reliable statistical analyses, it appears that the hydrophilic amino acids are slightly more preferable on the surface for connecting with another chain, while the hydrophobic amino acids are slightly more preferable on the surface facing the interior of their own chain.

(6) The N-cap and C-cap of the α -helix represented, respectively, by the codes $c(T_1) = c(T_5) = \text{FHABCDEG}$ but $c(T_8) \neq \text{FHABCDEG}$, and $c(T_1) = c(T_8) = \text{FHABCDEG}$ but $c(T_5) \neq \text{FHABCDEG}$, are examined. The frequencies of amino acid occurrences have much more particular patterns compared with those for the interior of the α -helix. Unfortunately, the amount of data is too small to perform the statistical analyses with confidence.

References

- [1] P. Y. Chou and G. D. Fasman, *Biochemistry*, **13**, 211-221 (1974).
- [2] M. Levitt, *Biochemistry*, **17**, 4277-4285 (1978).
- [3] H. Wako, N. Saito and H. A. Scheraga, *J. Protein Chem.*, **2**, 221-249 (1983).
- [4] R. W. Williams, A. Chang, D. Juretic and S. Loughran, *Biochim. Biophys. Acta.*, **916**, 200-204 (1987).
- [5] K.-H. Altmann, J. Wójcik, M. Vásquez and H. A. Scheraga, *Biopolymers.*, **30**, 107-120 (1990).
- [6] T. L. Blundell and Z.-Y. Zhu, *Biophys. Chem.*, **55**, 167-184 (1995).
- [7] A. Chakrabarty and R. L. Baldwin, *Adv. Protein Chem.*, **46**, 141-176 (1995).
- [8] S. Kumar and M. Bansal, *Proteins*, **31**, 460-476 (1998).
- [9] S. Kumar and M. Bansal, *Biophys. J.*, **75**, 1935-1944 (1998).
- [10] T. E. Creighton, *Proteins: Structures and Molecular Properties*. 2nd ed. W. H. Freeman and Company, New York, (1983).
- [11] L. G. Presta and G. D. Rose, *Science*, **240**, 1632-1641 (1988).
- [12] R. Aurora, R. Srinivasan and G. D. Rose, *Science*, **264**, 1126-1130 (1994).
- [13] R. Aurora and G. D. Rose, *Protein Sci.*, **7**, 21-38 (1998).
- [14] Z.-Y. Zhu and T. L. Blundell, *J. Mol. Biol.*, **260**, 261-276 (1996).
- [15] M. Schiffer and A. B. Edmundson, *Biophys. J.*, **7**, 121-135 (1967).
- [16] H. Wako and T. Yamato, *Protein Eng.*, **11**, 981-990 (1998).
- [17] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, M. D. Brice, J. R. Rodgers, O.

- Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, **112**, 535-542 (1977).
- [18] U. Hobohm, M. Scharf, R. Schneider and C. Sander, *Protein Sci.*, **1**, 409-417 (1992).
- [19] U. Hobohm and C. Sander, *Protein Sci.*, **3**, 522-524 (1994).
- [20] J. U. Bowie, N. D. Clarke, C. O. Pabo and T. Sauer, *Proteins*, **7**, 257-264 (1990).
- [21] J. U. Bowie, R. Luthy and D. Eisenberg, *Science*, **253**, 164-170 (1991).
- [22] M. J. Sippl, *J. Mol. Biol.*, **213**, 859-883 (1990).
- [23] D. T. Jones, W.R. Taylor and J. M. Thornton, *Nature*, **358**, 86-89 (1992).
- [24] M. Ota and K. Nishikawa, *Protein Eng.*, **10**, 339-351 (1997).
- [25] J. Overington, M. S. Johnson, A. Sali and T. L. Blundell, *Proc. Roy. Soc. London, Ser B.* **241**, 132-145 (1990).
- [26] J. Overington, D. Donnelly, M.S. Johnson, A. Sali and T. L. Blundell, *Protein Sci.*, **1**, 216-226 (1992).
- [27] H. Wako and T. L. Blundell, *J. Mol. Biol.*, **238**, 682-692 (1994).
- [28] H. Wako and T. L. Blundell, *J. Mol. Biol.*, **238**, 693-708 (1994).