Biophysical Chemistry, 47 (1993) 163–178
Elsevier Science Publishers B.V., Amsterdam

BIOCHE 01763

Conformational classification of short backbone fragments in globular proteins and its use for coding backbone conformations

Katsutoshi Takahashi and Nobuhiro Gō *

Department of Chemistry, Faculty of Science, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto, 606 (Japan)

(Received 22 December 1992; accepted in revised form 15 February 1993)

Abstract

An objective and systematic method of coding backbone conformations of proteins is developed. For this purpose polypeptide backbone is regarded to consist of two types of overlapping structural units, viz. a (ϕ, ψ) fragment (C_i^α–C′O–N–C_{i+1}^α–C′O–N–C_{i+2}^α), and a (ψ, ϕ) fragment (N–C_i^α–C′O–N–C_{i+1}^α–C′). By means of the principal component analysis, these (ϕ, ψ) and (ψ, ϕ) fragments are found to form five and six distinct clusters in respective high-dimensional conformational space with only a small number of exceptional unclassified fragments. Boundaries of clusters are defined by multi-dimensional ellipsoids. This classification of conformations of short backbone fragments is used to code protein backbone three-dimensional structures. In particular, conformations of peptide fragments consisting of four or six residues are coded and analyzed in detail. This structural code allows us to distinguish such features of protein backbone structures that are not retained in the usual secondary structural representation based on patterns of backbone hydrogen bonds, e.g., various types of four-residue turns. Similarity index, based on the number of different one-letter codes in a pair of structural codes, is a powerful measure of conformational similarity. When combined with another similarity measure, atomic root-mean-square distance, accurate similarity between a pair of conformations of fragments can be detected.

Keywords: Protein backbone conformation; Classification of three-dimensional structures; Principal component analysis; Structural code; Four-turn structures

1. Introduction

We are seeing a steady increase of a number of proteins whose three-dimensional structures have been determined experimentally. Together with the very rapid increase of the sequence information, we came to realize that a possible number of protein families may be rather limited [1]. This view leads us to expect that comparing protein three-dimensional structures would become even more important than ever.

Comparison of protein three-dimensional structures is deeply related with their coding. The

* *Correspondence to:* Professor Nobuhiro Gō, Department of Chemistry, Faculty of Science, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto, 606 Japan, Tel. 81-75-753-4017, Fax 81-75-711-6083.

purpose of comparison is, in a general sense, to find common features among a set or sub set of proteins. To each of such common features a code can be assigned. Conversely any codes that have been used for three-dimensional description are related to such common features. In this sense good coding is the purpose of comparison. However, at the same time, a good set of codes can be used to compare three-dimensional protein structures, not directly at the level of atomic cartesian coordinates but at the level of a set of codes. Because of the expected increasing importance of comparison of three-dimensional protein structures, we are interested in developing an objective coding system.

Proteins have a hierarchy of structure. In this paper, we are interested, roughly speaking, in the level of secondary structures. The most widely used definition of secondary structures is the one proposed by Kabsch and Sander [2]. This method assigns one of four secondary structural states (H, E, T and C) to each residue accoding to backbone hydrogen bonding patterns. Even though this coding is very useful for description of certain features of secondary structures, it is rather insensitive to variations of peptide backbone conformations.

Other often used systems of coding of local backbone structures are based on division of the two-dimensional (ϕ_i, ψ_i) space into several regions [3–7]. Usually division is made by straight boundaries, which inevitably introduce some arbitrariness.

In this paper, a novel approach is proposed to classify and code protein backbone structures based on coordinates of backbone atoms (N, C^{α}, C′ and O). This method removes the arbitrariness inherent in the coding system based on the division of the (ϕ, ψ) space, and supplements the Kabsch–Sander description of the secondary structures.

First, a mathematical background used for the development of the code is described. Second, this method is applied to actually develop a new system of coding. Third, usefulness of this coding system is demonstrated by applying it for description of various interesting features of protein structures.

2. Method

2.1. Preparation of a database of backbone fragments

We consider two types of backbone fragments shown in Figs. 1 and 2, respectively. When the peptide non-planarity is neglected, we can describe conformations of these fragments in terms of a pair of backbone dihedral angles (ϕ_i, ψ_i) or (ψ_i, ϕ_{i+1}). Therefore we call these fragments (ϕ, ψ) fragments (Fig. 1) and (ψ, ϕ) fragments (Fig. 2). Note that a (ϕ, ψ) fragment contains nine backbone atoms and two peptide planes and a (ψ, ϕ) fragment contains seven backbone atoms and one peptide plane.

In order to study the distribution of conformations of these fragments in their respective conformational space, we first construct a database of three-dimensional backbone structures of the fragments. Such data are taken from atomic coordinates in the Brookhaven Protein Data Bank [8]. Only those proteins whose resolution is 2.0 Å or better are considered. In case there are more than one entry for the same protein either because of existence of more than one protein in an asymmetric unit or because of different experimental or refinement procedures, only one set of data (with the best quality) is taken. When there are data for a family of evolutionary homologous proteins, one representative protein is taken. From these considerations we have selected 76 proteins listed in Table 1. The numbers of (ϕ, ψ) and (ψ, ϕ) fragments thus extracted are 12492 and 12578, respectively.

2.2. Distribution of conformations of the fragments in the conformational space

We now express the distribution of conformations of each type of fragments as a distribution

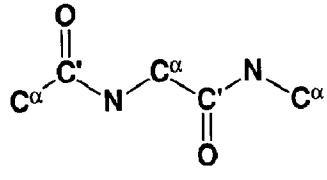

Fig. 1. Definition of the (ϕ, ψ) fragment.

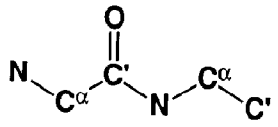

Fig. 2. Definition of the (ψ, ϕ) fragment.

of points in a high dimensional conformational space corresponding to each type of fragments. Any conformation of a fragment can be expressed at the atomic resolution by a point in a $(3N-6)$ dimensional space, where $3N$ is the number of Cartesian components of backbone atoms in the fragment (with N being the number of atoms) and 6 is the number of external degrees of freedom to remove the translational and rotational degrees of freedom to compare conformations. This removal is done by bringing each conformation to have such position and orientation in space that fits best with a certain reference conformation.

Best fitting is done by minimizing analytically [9] a distance between two conformations. Here the distance is defined by the square root of the mass weighted mean square of distances between corresponding atoms in the two conformations.

The reference conformation, against which each of all other conformations is best-fitted, is chosen from among all conformations of the fragments so that the sum of distances to all non-reference conformations is the minimum.

We should note one point in the above procedure of the best-fitting. We have translated and rotated each of conformations, A,B,..., to have a minimum distance from the reference conformation R. From the resulting Cartesian coordinates we can calculate a distance between, for instance, A and B. Let us designate the resulting distance by d_R(A,B), because this distance is determined in reference to conformation R. But we can also define a distance between the two conformations A and B by best-fitting B directly to A, or, in other words, by minimizing the distance between A and B. The resulting minimum distance, which we shall designate by d_D(A,B), is in general smaller than the distance calculated above, i.e., that calculated from the Cartesian coordinates A and B, when both of them are best-fitted to R.

The above difference between two definitions of distance between two conformations A and B is related to a mathematical structure of the conformational space we are dealing with. The conformational space is a metric space in the sense that distance is defined between points in the space. The distance d_D(A,B) is better from the point of view that it does not depend on a reference conformation. However, the space with distance defined by d_D(A,B) is a non-Euclidian space, whereas that defined by d_R(A,B) is a Euclidian space. The difference between the two distances reflects the curvature of the non-Euclidian space. In this paper we employ the distance defined by d_R(A,B) in order to avoid mathematical complexities associated with the non-Euclidian nature of the space when the other choice is employed.

Thus, we express various conformations of each type of fragments as points in a metric space where the distance is defined by d_R(A,B). A backbone conformation of a fragment can be expressed in this multi-dimensional space as a point whose coordinates are $x_i = m_i^{1/2} r_i / M^{1/2}$, $(i = 1, 2, \ldots, 3n_a)$, where n_a is the number of backbone atoms in the fragment, and m_{3j-2}, m_{3j-1}, m_{3j} and r_{3j-2}, r_{3j-1}, r_{3j} are mass and Cartesian coordinates of jth atom in the fragment, respectively, and M is the sum of mass of backbone atoms in the fragment. This conformational space is $3n_a$ dimensional. In this space conformational points are not distributed isotrop-

Table 1

Entries of proteins in our three-dimensional structural database [a]

1ACX	1BP2	1CRN	1CTF	1ECD	1FB4	1FDX	1FX1
1GCR	1HDS	1H1P	1HMQ	1INS	1LZ1	1LZT	1MLT
1NTP	1NXB	1OVO	1PCY	1PPD	1PPT	1REI	1RN3
1RNT	1RSM	1SGC	1SN3	1TGN	1TPA	1TPP	1UBQ
2ACT	2ALP	2APP	2APR	2AZA	2B5C	2CCY	2CDV
2CGA	2CNA	2CPP	2CTS	2CYP	2EBX	2FD1	2GRS
2LH4	2LHB	2LYM	2LZM	2MHB	2MHR	2OVO	2PAB
2PTC	2PTN	2RHE	2SGA	2SNS	3BCL	3C2C	3DFR
3EST	3HHB	3MBN	3RXN	451C	4DFR	4FXN	4PTI
5CPA	5RXN	6RSA	9PAP				

[a] Entry name in Brookhaven Protein Data Bank [8] is given.

ically, but, in fact, is distributed widely along a rather small number of dimensions and narrowly along the others. To find such dimensions with large distributions, we carry out the principal component analysis described below.

Let us assume that we have n conformations in our data set. We denote ith coordinate of kth conformation by x_{ki}. At first we choose the origin of the multi-dimensional conformational space to coincide with the center of distribution. This can be achieved by subtracting from x_{ki} its average over n conformations. Now we redefine x_{ki} as a quantity whose average is subtracted. Then, we introduce a new coordinate system characterized by a set of unit coordinate vectors (principal coordinate vectors), $f_1, f_2, \ldots$, where $f_1, f_2, \ldots$, are unit vectors of the direction in the $3n_a$ dimensional conformational space along which the n points are the most widely distributed, the second most widely distributed, and so forth. Let ith component of lth unit vector f_l be f_{il}. In this new coordinate system each point is represented by a set of new coordinates given by

$$y_{kl} = \sum_i x_{ki} f_{il}. \tag{1}$$

The first principal coordinate vector f_1 is determined from the condition that the following quantity, which means the mean square distribution of conformational points along this direction, be maximum.

$$\frac{1}{n} \sum_k y_{kl}^2 = \sum_{i,i'} C_{ii'} f_{il} f_{i'l} \tag{2}$$

where $l = 1$ (for the first principal coordinate), and $C_{ii'}$ are elements of the variance–covariance matrix $\boldsymbol{C}$ defined by

$$C_{ii'} = \frac{1}{n} \sum_k x_{ki} x_{ki'}. \tag{3}$$

Because f_l are unit vectors, they satisfy the following equation.

$$\sum_i f_{il}^2 = 1 \tag{4}$$

The vectors of solution of the extremum problem of eq. (2) under the constraint of eq. (4) satisfies the following eigenvalue problem.

$$\sum_{i'} C_{ii'} f_{i'l} = \lambda_l f_{il} \tag{5}$$

Because the unit vectors f_l are given as eigenvectors of the eigenvalue problem of eq. (5), they are orthogonal to each other. From this orthogonality and eqs. (2) and (5), we can see that the eigenvalue λ_l means the mean square distribution of n points along the lth principal coordinate axis. Because eq. (5) is the secular equation for a $3n_a \times 3n_a$ positive semi-definite symmetric matrix, there are $3n_a$ eigenvalues. Corresponding to the operation of best-fitting of fragments to a reference conformation, we should have six zero eigenvalues among the $3n_a$ eigenvalues. The total mean square distribution is given by a sum of λ_l over l, or by $\mathrm{tr}(C)$. Therefore,

$$(\text{f.m.s.d.})_l = \lambda_l / \mathrm{tr}(C) \tag{6}$$

is fractional mean square distribution (f.m.s.d.) of points along lth principal coordinate axis.

In the following we will see that the distribution of conformational points can be quite well represented in a subspace of several dimension corresponding to the first several principal coordinate axes instead of the very large $3n_a$ dimensional space. This enables us to "visualize" the distribution of points.

3. Results

3.1. Distribution of conformations of (ϕ, ψ) and (ψ, ϕ) fragments and their classification

There are nine backbone atoms in a (ϕ, ψ) fragment (Fig. 1). Therefore this fragment has 27 degrees of freedom. Consequently, we should and in fact did obtain 21 non-zero and 6 zero eigenvalues as a result of the principal component analysis. However, conformational differences among various fragments can be expressed by a small number of principal components. Actually,

92% of the entire distribution can be expressed by only four principal components.

In the (ϕ, ψ) fragments there are two peptide groups, each of which is nearly planar. Because of the planarity, there are only two easily rotatable dihedral angles in these fragments. Therefore, conformations of these fragments may be understood as distributed in a nearly two-dimensional space corresponding to the dihedral angles (ϕ, ψ). However, this space has the topology of torus and non-Euclidian. Such space can be embedded only into higher dimensional Euclidian space. This is the reason why the distribution cannot be expressed well by two principal components.

We now try to visualize the distribution of the fragments. Figure 3(a) shows the projection of conformational points onto the two-dimensional space corresponding to the first and the second principal components. The distribution shown in Fig. 3(a) accounts for 73.1% of the entire distribution. It appears in Fig. 3(a) that the conformational points are distributed in three clusters, designated as α, β and γ. To understand the meaning of these clusters we calculate projection of the (ϕ_i, ψ_i) plane (each point on which corresponds to a conformation of the fragment with standard bond lengths and bond angles) onto the two-dimensional space of the first two principal components and show in Fig. 3(b). Figure 3(c) shows the distribution of Fig. 3(a) as superposed to the projected (ϕ, ψ) plane in Fig. 3(b). We can see in Fig. 3(c) that the conformational points of the fragments are distributed within a thin layer from the two-dimensional hypersurface of the (ϕ, ψ) plane. The thickness of the layer should correspond to out-of-plane deformations of peptide planes and deviations of bond lengths and bond angles from standard values.

Examination of the projection of the (ϕ, ψ) plane in Fig. 3 suggests that the three clusters correspond to the well-known three clusters in the α-helix, β-strand and left-handed α-helix regions, respectively. However, each of these clusters appearing to form a cluster in the projected two dimensional space, can still be distributed in two or more clusters in the full higher dimensional space. Examination of this possibility is done as follows. At first all conformational points are classified into one of the three clusters by drawing straight lines in the two-dimensional distribution in Fig. 3(a). Then, the principal component analysis is carried out again for each of the clusters identified above. In this analysis the distance between a pair of fragments A and B, d_R(A,B), is defined in reference to a new fragment R, which is located near the center of the distribution of each cluster. Among the three clusters, β is found to have again one continuous cluster in the new two-dimensional projection. Cluster α is found to be consisting of one major cluster (α_1), one minor cluster (α_2) and scattered points. Cluster γ is also found to be consisting of one major cluster (γ_1), one minor cluster (γ_2) and scattered points. When we again apply the principal component analysis to the four new clusters, α_1, α_2, γ_1 and γ_2 (again defined by straight boundaries in the two-dimensional distribution), all of them are found to have no further subclusters.

Similar procedure has been taken for the (ψ, ϕ) fragments, too. There are seven backbone atoms in this type of fragment. Five atoms from C^α to the next C^α form one peptide plane. Positions of N and C′ with respect to the peptide plane is determined by the dihedral angles ψ_i and ϕ_{i+1}, respectively. The first four principal components account for 95% of the entire distribution of the conformational points. When the distribution is projected onto the two-dimensional space corresponding to the first two principal components, four clusters were observed. Projection of distribution in each cluster onto the two-dimensional space corresponding to the first two principal components of each cluster reveals that two clusters consist of two subclusters and of three subclusters, respectively. Thus, the (ψ, ϕ) fragments are found to occur in seven clusters, A (Alpha), B_1 ($Beta_1$), B_2 ($Beta_2$), Γ_1, Γ_2, Γ_3 and Δ, and in other scattered points.

So far the five clusters of the (ϕ, ψ) fragments have been defined by straight boundaries in a two-dimensional distribution. But, drawing the straight lines involves some arbitrariness. Therefore such a definition of the clusters inevitably suffers from the arbitrariness. In order to remove such an arbitrariness, we now redefine each clus-

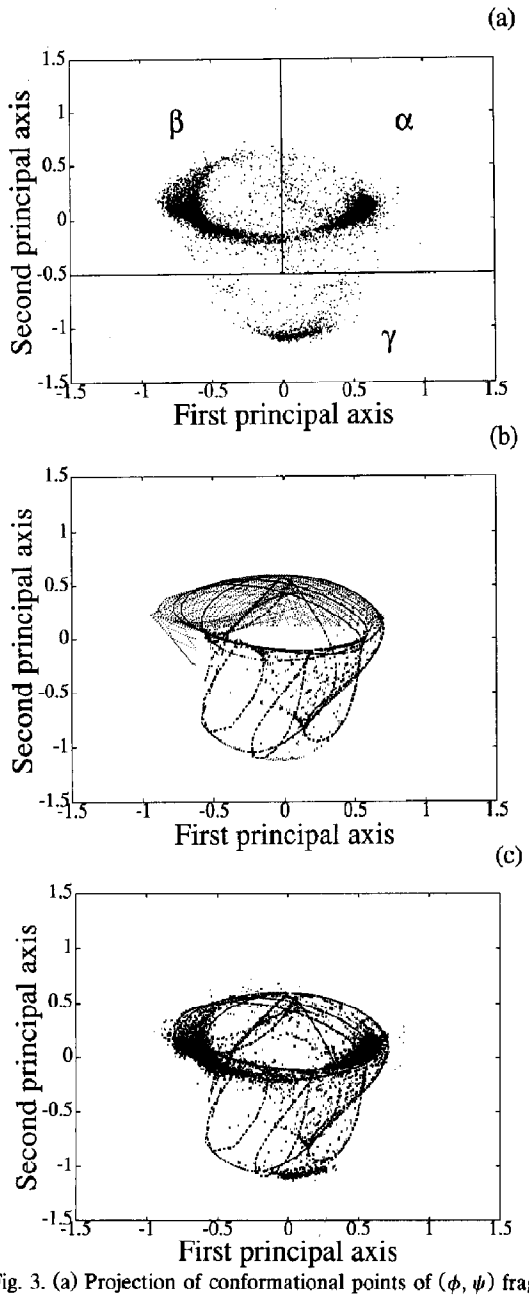

Fig. 3. (a) Projection of conformational points of (ϕ, ψ) fragments onto the two-dimensional space corresponding to the first two principal components. (b) Projection of the (ϕ_i, ψ_i) plane onto the same space. Trajectories of constant values of every ten degrees of ϕ and ψ are shown by thin dotted lines. Standard (ϕ, ψ) values of the right-handed α-helix, the left-handed α-helix, the parallel and anti-parallel β-strand cited in an IUPAC-IUB document [10] are shown by *, +, O and X. Trajectories of constant values of ϕ or ψ going through these points are indicated by thick solid or broken lines, respectively. (c) Superposition of the distribution in (a) to the projected (ϕ, ψ) plane in (b).

ter by an ellipsoidal boundary. The dimension of the ellipsoid is taken equal to the number of principal components, which are sufficient to account for 90% of the distribution of the conformational points in each cluster. This number, the dimension of the ellipsoid, is listed in Table 2. Also the fraction of the distribution, that can be accounted for by this number of principal components, is given in Table 2. The size of the ellipsoid is taken to be R times the standard deviation along each principal component. The values of R for five clusters are determined by requiring that overlaps between ellipsoids of different clusters be minimum and the number of conformational points covered by the ellipsoids be maximum. The values of R thus determined are also listed in Table 2. Similar procedure is taken for the seven clusters in the (ψ, ϕ) fragments. During this procedure clusters B_2 and Δ are found to have a very large overlap. Therefore, we merged them to redefine new Δ. The results are also listed in Table 2. Usage of ellipsoids of several dimensions to define clusters in the above means that a fragment is judged to belong to a cluster, if it is contained in an ellipsoid, no matter where

Table 2

One-letter code and boundary ellipsoids to characterize each cluster

Cluster	One-letter code	Number of principal components	Fractional distribution [a] (%)	R [b]
α_1	a	6	92.0	5.5
α_2	b	6	96.6	3.1
β	c	4	92.6	3.5
γ_1	d	4	90.3	4.5
γ_2	e	4	92.4	2.2
A	A	5	91.6	4.0
B_1	B	4	93.5	4.0
Γ_1	C	4	90.5	4.0
Γ_2	D	5	94.2	4.0
Γ_3	E	5	91.8	2.1
Δ	F	5	92.2	4.0

[a] Fraction of distribution in percent of conformational points in each cluster that can be accounted for by the first several principal components whose number is listed as the number of principal components.

[b] The size of the ellipsoid is taken to be R times the standard deviation along each principal component.

the conformational point is located in the remaining large number of dimensions not considered in the definition of the ellipsoid.

The conformational clusters obtained above can be used to code protein backbone structures. In the following, we will use the one-letter code listed in Table 2 for each cluster of (ϕ, ψ) and (ψ, ϕ) fragments. If a (ϕ, ψ) or a (ψ, ϕ) fragment cannot be assigned as a member of any of these clusters, then a one-letter code 'o' or 'O', respectively, is assigned to it. Any conformation of long peptide backbone fragment can be expressed by a string of these one-letter codes for (ϕ, ψ) and (ψ, ϕ) fragments contained in it.

Figures 4(a) and 4(b) show how conformational points are distributed in the vicinity of each cluster. In all cases of five clusters of (ϕ, ψ) fragments and six clusters of (ψ, ϕ) fragments, a clear peak of distribution is seen. This indicates that the boundaries of clusters of distribution of conformational points in the multi-dimensional space can be well defined by the several dimensional ellipsoids. The values of R listed in Table 2 are chosen to include the first peaks. The determination of the clusters and their boundaries have been done by observing the projections of the distributions onto the two-dimensional space corresponding to the first two principal components. Figure 4 indicates that the multi-dimensional distribution can be well observed by its two dimensional projection.

Positions and widths of the peaks of the frequency distributions are similar to those of the Gaussian distribution. In some cases peaks are located at smaller values of the scaled distance r. Thus, the conformational points have basically Gaussian-like distributions, but in some cases distribution is more weighted at smaller values of r. The distributions of clusters 'e' and 'E' appear to have some overlaps with distributions of other clusters. The values of R have been chosen to minimize the overlaps, which forced the values of R for the two clusters to be slightly smaller than the ones appearing proper from the Gaussian-like distribution.

Now, given a conformation of a fragment, we

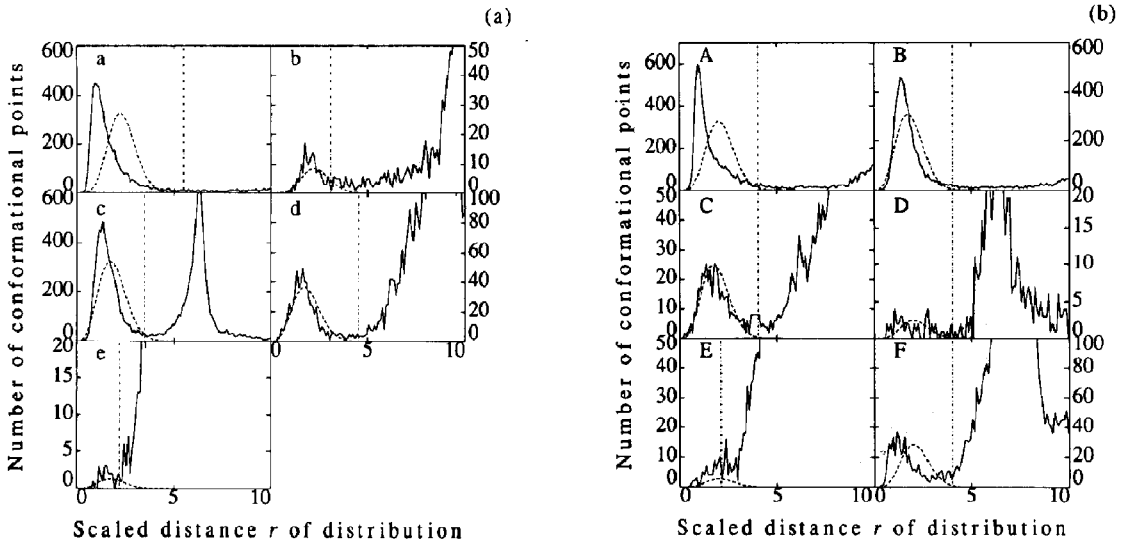

Fig. 4. Frequency distribution of conformational points in the vicinity of each cluster. Position of a conformational point in the distribution of one cluster is measured by the scaled distance r, which means that the conformational point is located at a position which is r times the standard deviation of the distribution of the cluster. Ordinate is the number of conformational points in each 0.1 interval of r. Vertical dot-dashed lines indicate the value of the scale factor adopted to determine the boundaries of the multi-dimensional ellipsoids. Broken lines show theoretical frequency distribution function, when the cluster is distributed according to the Gaussian distribution. (a) Frequency distributions of (ϕ, ψ) fragments. (b) Frequency distributions of (ψ, ϕ) fragments.

can judge if it belongs to one of the clusters by checking if its conformational point exists within one of the several dimensional ellipsoids. Computer program COSP (COdes for Structure in Protein), can do this classification of conformation of fragments and, coded in FORTRAN, is deposited in Genome-net anonymous-FTP server (Internet-name: ftp.genome.ad.jp). Each of the boundary ellipsoids is defined in terms of the coordinates of the center of the ellipsoid, and vectors of the principal axes, the standard deviation of distribution along each principal axis, and the value of R. Values of these quantities are given explicitly in COSP.

3.2. Characterization of each cluster

Classification of the 12492 (ϕ, ψ) fragments by the boundary ellipsoids is shown in Table 3. Clusters a and c are very populated, d and b intermediate, and e contains but a small number of fragments. Out of the 12492 fragments, 11985 fragments (95.9%) are classified into the five classes.

Main chain dihedral angles in the fragments contained in each class are plotted in Fig. 5. The five clusters identified in the multi-dimensional conformational space are seen also well separated in the familiar (ϕ, ψ) plane in Fig. 5(a). From the locations in this familiar plane clusters a and c are identified, respectively, as the right-handed α-helical structure and the extended β-strand structure. Wilmot and Thornton [6] proposed to divide the β region in the (ϕ, ψ) plane into two parts, β_E and β_P. However, the result of our principal component analysis indicates that,

Table 3

Classification of 12492 (ϕ, ψ) fragments by the boundary ellipsoids

Letter	a	b	c	d	e	o
a	5628	–	4	–	–	
b		150 (139)	–	–	–	
c			5554	2	–	
d				625	–	
e					22 (21)	
o						507 (519)

Diagonal elements show the number of fragments contained in only one ellipsoid. The number in the parentheses is for re-defined cluster by adding one more criterion that the central residue be Gly in case of cluster b and Pro in case of cluster e. Off-diagonal elements show the number of fragments contained in two ellipsoids. The letter o stands for *others*.

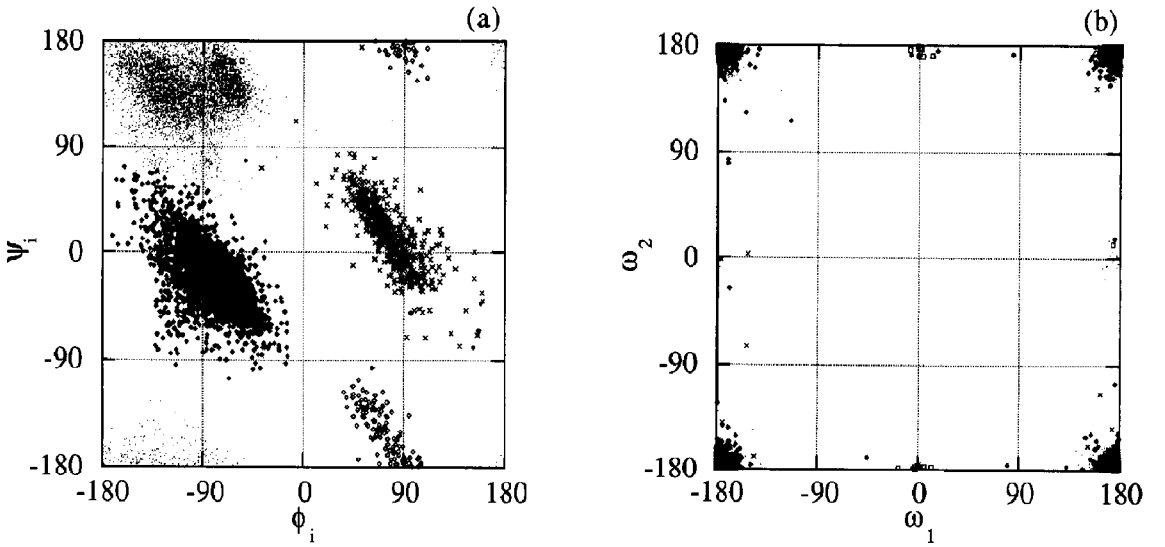

Fig. 5. Main chain dihedral angles in fragments contained in each class of (ϕ, ψ) fragments: (♦) a, (◇) b, (●) c, (×) d and (□) e. (a) Plot of (ϕ_i, ψ_i). (b) Plot of (ω_1, ω_2), where ω_1 and ω_2 are dihedral angles ω of N-terminal and C-terminal peptide planes, respectively.

from the point of view of the three-dimensional structures of the fragments, such division is not necessary.

Fragments contained in cluster d have a pair of dihedral angles characteristic of a left-handed α-helical structure. It is known that only Gly is allowed in the left-handed α-helical region, when standard geometry and a non-deformable peptide plane is assumed. In fact the central residue of 61.2% of the 625 fragements in cluster d is found to be Gly. However, the remaining 38.8% are non-Gly residues. Among the non-Gly residues Asn is most abundant, accounting for 13.1% of the 625 fragments. When distributions of these Gly and non-Gly residues in cluster d are plotted in the (ϕ, ψ) plane, the latter is found to be centered towards the left-upper corner of the left-handed α-helical region. It has been pointed out [3,11–13] that the interaction between polar sidechain and polar amide backbone atoms allow an Asn residue to take this left-handed α-helical backbone conformation.

Cluster b occurs in a region of the (ϕ, ψ) plane which is allowed only for Gly. In fact, the central residue of 92.7% of the fragments in this cluster is Gly. Examination of remaining cases (11 out of 150 fragments) shows that in most of these remaining cases there are severe steric overlaps, indicating that these fragments are observed in this region of the (ϕ, ψ) plane due to experimental error. Therefore we redefine the b cluster by adding one more criterion that the central residue be Gly.

Fragments in cluster e are distributed in the (ϕ, ψ) plane in the same region as for cluster c. However, the value of peptide ω of the peptide plane on the N-terminal side is in the region corresponding to *cis*-peptide plane. In fact, the central residue of 95.5% (21 out of 22) of the fragments in this cluster is Pro. Therefore, this cluster e corresponds the local conformation characteristic to *cis*-Pro. We redefine cluster e by adding one more criterion that the central residue be Pro.

Fragments in which the C-terminal side peptide plane assumes *cis*-conformations are contained in cluster c. This means that most of residues preceding *cis*-Pro assumes an extended structure, reflecting the helix-breaking nature of Pro.

Classification of 12578 (ψ, ϕ) fragments by the boundary ellipsoids is shown in Table 4. There are two clusters, A and B, with large population, two clusters with intermediate population, C and F, and two clusters with small population, D and E. E has some overlap with A and B, indicating

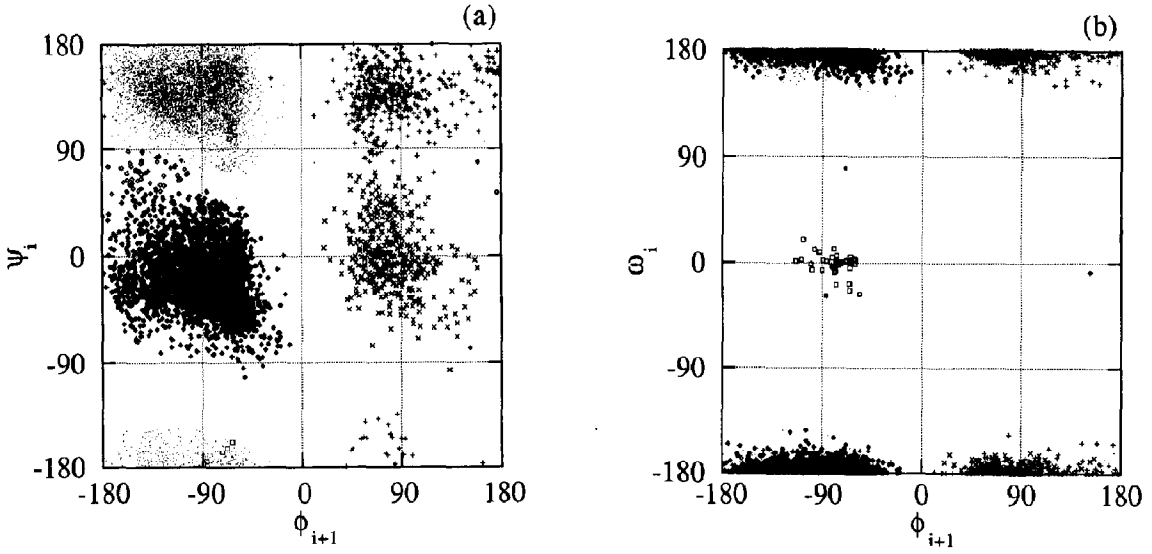

Fig. 6. Mainchain dihedral angles in fragments contained in each class of (ψ, ϕ) fragments: (♦) A, (●) B, (×) C, (□) D, (◇) E and (+) F. (a) Plot of (ψ_i, ϕ_{i+1}). (b) Plot of (ω_i, ϕ_{i+1}).

Table 4

Classification of 12578 (ψ, ϕ) fragments by the boundary ellipsoids

Letter	A	B	C	D	E	F	O
A	5665	–	–	–	20	–	
B		5288	–	–	5	11	
C			428	1	–	1	
D				43 (39)	–	–	
E					46	–	
F						494	
O							576 (580)

Diagonal elements show the number of fragments contained in only one ellipsoid. The number in the parenthesis is for re-defined cluster D by adding one more criterion that the C-terminal side residue be Pro. Off-diagonal elements show the number of fragments contained in two ellipsoids. The letter O stands for *others*.

its intermediate structural character between A and B. Out of the 12578 fragments, 12002 fragments (95.4%) are classified into the six classes.

Plot of mainchain dihedral angles in the (ψ, ϕ) fragments are given in Fig. 6. These plots indicate that A and B are structures appearing in the α-helix and the extended β-strand.

Fragments in cluster D are distributed in the same region as that of B in Fig. 6(a). However, Fig. 6(b) indicates that they have *cis*-peptide planes. In fact C-terminal residue of 90.9% (40 out of 44) of the fragments in D is found to be Pro. Because the remaining four are likely to be contamination due to experimental error, we redefine this cluster by adding one more criterion that its C-terminal residue be Pro. It is noted that the occurrence of tyrosyl residue in the N-terminal side of this fragment is remarkably high (18.2%) than the average occurrence in the whole database. Conformations containing *cis*-Pro on the C-terminal side are also found among fragments assigned as O.

Fragments in cluster C are distributed in a region which is the mirror image of the region of distribution of A with respect to the ϕ-axis. The C-terminal residues are 70.0% Gly and 7.4% Asn. No strong residue type preference is observed in the N-terminal side with Asn appearing relatively frequently (11.5%).

Fragments in cluster F are distributed in a region which is the mirror image of the region of distribution of B with respect to the ϕ-axis. The C-terminal residues are 66.2% Gly and 10.5% Asn. No strong residue type preference is observed in the N-terminal side with Ser appearing relatively frequently (14.0%). Unlike in C, Asn appears in the N-terminal side only 2.6%.

Fragments in cluster E are distributed in a region between those occupied by A and B. Occurrence of amino acid types are rather unusual. Popular residues at the C-terminal side are Ile (16.9%), Ser (14.1%) and Asn (9.9%). Those at the N-ternimal side are Asp (21.1%), Asn (15.5%) and Gly (12.7%).

3.3. A rule for a string of one-letter codes

Table 5 summarizes numbers of occurrence of a particular cluster of (ψ, ϕ) fragments following a particular cluster of (ϕ, ψ) fragments in the three-dimensional structures of the 76 proteins examined. Table 6 is for (ϕ, ψ) fragments following (ψ, ϕ) fragments. Cases with a small occurrence number are likely due to experimental errors. When we neglect them, there appears to exist rather strict rules of limitation of occurrence. These limitations come (a) from limited ranges shown in Figs. 3 and 4 of values of dihedral angles common in consecutive (ϕ, ψ) and (ψ, ϕ) fragments, and (b) from limitations of the types of amino acid residues in some of the clusters.

By combining two Tables 5 and 6, we can deduce the rule for occurrence of types of (ψ, ϕ)

Table 5

Numbers of occurrence of a particular cluster of (ψ, ϕ) fragments following a particular cluster of (ϕ, ψ) fragments

(ϕ, ψ) fragments	Number of occurrence (ψ, ϕ) fragments						
	A	B	C	D	E	F	O
a	4426		266		2		149
b		55				6	53
c		3801	3	15	21	329	64
d	342	5	62		5		54
e		18		1			
o	52	141	14	20	5	30	176

fragments between a given pair of (ϕ, ψ) fragments. For example, A, C and E occur after a, and A, B and E occur before a. Therefore, A and E an occur between a and a to make two triplets, aAa and aEa. From the prepared database we calculated number of occurrences of all types of triplets and found that aAa occurs 3724 times and aEa occurs just once. From the rarity of aEa, we suspect that it may be a noise occurring from an experimental error. Examination of all other cases indicates that, between a given pair of (ϕ, ψ) fragments, either no or essentially one type of (ψ, ϕ) fragment can occur except between c and c. In the latter case two triplets, cBc and cEc, are possible. This reflects the character of the (ψ, ϕ) fragment E, which is intermediate between A and B. The fact that the type of central (ψ, ϕ) fragment is determined almost uniquely from a pair of preceding and following (ϕ, ψ) fragments is natural, because, given the types of two (ϕ, ψ) fragments, ranges of values of two central dihedral angles are well determined.

3.4. Helices and extended structures

We now give in Table 7 structural codes for peptide backbone fragments consisting of four amino acid residues assuming the α-helix, the G-helix or the extended structure. A four-residue fragment contains three (ψ, ϕ) fragments and two (ϕ, ψ) fragments. Therefore, its conformation is described by a string of five one-letter codes.

Table 6

Numbers of occurrence of a particular cluster of (ϕ, ψ) fragments following a particular cluster of (ψ, ϕ)-fragments

(ψ, ϕ) fragments	Number of occurrence (ϕ, ψ) fragments					
	a	b	c	d	e	o
A	3774		960		1	56
B	879		3094	1		66
C		52		239		55
D					16	20
E	5		25			3
F		61	4	223		76
O	185	1	139	14	2	164

Table 7

Structural codes of standard secondary structures consisting of four residues [a]

Secondary structures	Codes
α-Helix	AaAaA
G-Helix	AaAaA
Parallel β-strand	BcBcB
Anti-parallel β-strand	BcBcB

[a] A four-residue fragment contains two (ϕ, ψ) fragments and three (ψ, ϕ) fragments. Thus, its backbone conformation is expressed by five one-letter codes, $X_1x_1X_2x_2X_3$, where X and x refer to a one-letter code for corresponding (ψ, ϕ) and (ϕ, ψ) fragment, respectively.

Among these structural codes of standard secondary structures, two secondary structures, the right-handed α-helix and G(3_{10})-helix, have the same structural code. Four residue fragments all assigned to α-helix by the Kabsch–Sander secondary structure coding have an average main-chain atomic (N, C^{α}, C′ and O) root-mean-square (r.m.s.) distance of 0.26 Å. The distance between a standard four-residue α-helix and G-helix is only 0.35 Å. Therefore, their local backbone structures are very similar, even though these two types of helices are characterized by different hydrogen bond networks. This similarity is reflected to the same structural code assigned for both of them.

The r.m.s. distance between the ideal parallel and anti-parallel β-strand [10] is only 0.29 Å, which is small compared with the conformational differences found among four residue fragments whose residues are all assigned to β-strand by the Kabsch–Sander coding. Thus, also in this case, the identical structural codes correctly reflect the conformational similarity.

3.5. Four-turns

Assigment of our one-letter codes to various four-turn structures is done as follows. At first we examine all backbone fragments consisting from C_i^{α} in $C_i^{\alpha} i_{+3}$ in the 76 proteins. When distance between C_i^{α} and C_{i+3}^{α} is less than 7 Å, we judge, following Lewis et al. [14], that the backbone fragment takes a four-turn structure, unless it is not in an α-helix. Backbone fragments consisting

Table 8
Various types of four-turns and their structural codes

Type	Code	Number of occurrence
I	aAa	773
	aOa	36
I′	dCd	50
II	cFd	190
II′	bOa	40
	bBa	3
VIII	aAc	139
$\gamma\alpha$	dAa	18
$\alpha\gamma$	aCd	11
$\beta_E\epsilon$	cFb	6
$\beta\alpha$	cOa	4
$\epsilon\beta_P$	bOc	4
VI	(c,o)D(e,o)	22

from C_i^α to C_{i+3}^α contains three peptide planes, and their conformations can be described by a three-letter code $x_1X_1x_2$. We assign three-letter codes for all fragments that form four-turns. Classification of these four-turn structures into standard types can be easily done by examination of values of the dihedral angles in them and of their three-dimensional structures by computer graphics. All cases of three-letter codes with three or more four-turns observed in the selected 76 proteins are classified into standard types and listed in Table 8. Note that structures coded as aAa appear also in α-helices. But, as mentioned above, they are excluded from Table 8. Note also that, even though Table 8 shows structural codes of backbone fragments recognized as forming turns, a fragment having one of the codes listed in Table 8 does not necessarily assume a turn conformation, i.e., the distance between C_i^α and C_{i+3}^α can be larger than 7 Å.

Turn-type classification by Wilmot and Thornton [6] is employed. In the classification by Wilmot and Thornton, they divided the β-region in the (ϕ, ψ) plane into two parts. However, as mentioned already, in our treatment based on the three-dimensional structures of fragments this region can be and is represented by only one cluster c.

Type I′ turn has quite frequently Gly at second and third residue position [15]. The central three-letter string dCd is consistent with this fact, because both clusters d and C have often have Gly residues.

Type II turn has the central three-letter code cFd. The C-terminal residue of cluster F is Gly rich. This is consistent with the well known fact that the third residue position in type II turn is often Gly.

The central three-letter code for type II′ turn is either bBa or bOa. Code b stands for a cluster which is allowed only for Gly as the central residue of the (ϕ, ψ) fragment, consistent with the fact this turn tends to have Gly at the second residue position. According to Tables 5 and 6, only B and E can occur between b and a. However, examination of Figs. 5(a) and 6(a) reveals that the ranges of distribution of the values of common diheral angle ψ in b and B are slightly different from each other. This slight difference causes the central one-letter code sometimes to be O rather than B.

Type VI turn coded as (c,o)D(e,o) is characterized by the presence of *cis*-Pro at the third C^α position. All four cases of cDe, cDo, oDe and oDo are observed.

3.6. *Frequent structures of four-residue and six-residue fragments*

All possible four-residue and six-residue overlapping fragments are taken from the selected 76 proteins. To each of the four-residue or six-residue fragments, a five-letter or nine-letter code is assigned, respectively. Tables 9 and 10 list frequent five-letter codes and nine-letter codes, respectively.

The most frequent five-letter code, AaAaA, and the most frequent nine-letter code, AaAaAaAaA, both correspond to the α-helical conformation. The second most frequent codes, BcBcB and BcBcBcBcB, correspond to the extended β-strand conformation. Other frequently observed codes, BaAaA (5th entry), BcBaA (4th), AaAcB (6th) and AcBcB (3rd), appear to correspond to boundary conformations between an α-helix and β-strand. Similar boundary conformations are also often found in the nine-letter codes.

It is remarkable that only 20 different five-letter codes and 29 different nine-letter codes account for 80% and 60%, respectively, of all of four-residue and six-residue fragments.

3.7. Structural similarity between short fragments having similar structural codes

A string of one-letter codes for a short peptide fragment gives a rather accurate description of local three-dimensional structures of (ϕ, ψ) and (ψ, ϕ) fragments in it. But, if a pair of fragments have similar or identical strings of one-letter codes, we may expect that their overall three-dimensional structures are also similar to each other. To see this point, the r.m.s. distances between pairs of fragments having the same structural code in Table 10 were calculated for all pairs. Their averages and standard deviations are also given in Table 10.

Table 9

Frequently observed structural codes for four-residue fragments [a]

Code	m_i	Fraction (%)	Cumulative fraction (%)
AaAaA	2891	28.47	28.47
BcBcB	1986	19.56	48.03
AcBcB	553	5.45	53.48
BcBaA	500	4.92	58.40
BaAaA	454	4.47	62.87
AaAcB	309	3.04	65.91
BaAcB	236	2.32	68.24
AcBaA	200	1.97	70.21
BcBcF	179	1.76	71.97
AaAaC	138	1.36	73.33
CdAcB	136	1.34	74.67
AaCdA	122	1.20	75.87
FdAcB	109	1.07	76.95
BcFdA	103	1.01	77.96
BaAaC	43	0.42	78.38
AcBcF	41	0.40	78.79
AaAcF	38	0.37	79.16
BcFdC	38	0.37	79.54
AcFdA	37	0.36	79.90
FdCdA	35	0.34	80.24

[a] Structural codes which contain neither o nor O are listed. The second column lists the number of fragments with the same five-letter code.

At first we see that the α-helical fragments have very sharply defined three-dimensional structures with an average of r.m.s. distances being 0.55 Å. In comparison, fragments with the extended conformation have relatively large structural variations with an average of 1.20 Å. Structural variations in fragments with most of the frequent structural codes are similar to the case of the extended conformations.

When the above analysis was done, we found that some six-residue fragments have very sharply defined three-dimensional structures. Their structural codes and degree of structural variations are also shown in Table 10. The fragments containing BcBcFdCdCdAcBcB should correspond to type I′ β-ladder. This structure is held with three intra-fragment hydrogen bonds, thus rendering the structure very rigid. The fragments with BaAaCdAcB corresponds to the so-called five-turn [15,16]. These structures are stabilized by two intra-fragment hydrogen bonds. The fragments with AaAaAcFdAcB, which probably have not been described before in the open literature, should be classified as type II $\alpha\beta$-loop, because connection of an α-helix to a β-strand occurs by a loop having a structure characteristic of a type II turn, i.e., cFd. Stereodrawings of these structures are given in Fig. 7.

So far we studied distance between a pair of fragments with identical structural codes. When the number of different one-letter codes increases, degree of structural dissimilarity should increase. Figure 8(a) shows distribution of r.m.s. distance between all possible pairs of six-residue fragments taken irrespective of their structural codes from the 76 proteins. Its breakdown into contributions from pairs of fragments with a given number of different one-letter structural codes is given in Fig. 8(b). Here one-letter structural codes o and O are always counted as a different code.

The first peak at about 0.3 Å in Fig. 8(a) comes from pairs with identical codes. Main contribution is from α-helical fragments, with a small amount of contributions from such fragments as listed at the bottom of Table 10. We see that there is also a peak at about 0.5 Å in contribu-

tions from pairs of fragments with one different one-letter code. There are also an appreciable number of pairs of fragments within 0.8 Å among pairs of fragments with two different one-letter codes. We found that these contributions occur only when o and/or O are involved.

3.8. Application to search for structurally very similar backbone fragments

Here we describe one application of the structural code. We have recently reported the frequent occurrence of common SARFs (Spatial

Table 10

Frequently observed structural codes for six-residue fragments [a] and some six-residue fragments with sharply defined structures

Code	m_i	Fraction (%)	Cumulative fraction (%)	μ [b]	σ [b]
AaAaAaAaA	2221	21.87	21.87	0.55	0.37
BcBcBcBcB	985	9.70	31.57	1.20	0.36
AcBcBcBcB	294	2.90	34.47	1.26	0.37
BaAaAaAaA	245	2.41	36.88	0.78	0.48
BcBcBcBaA	243	2.39	39.28	1.32	0.40
BcBaAaAaA	195	1.92	41.20	0.97	0.45
BcBcBaAaA	174	1.71	42.91	1.22	0.49
BcBcBcBcF	123	1.21	44.12	1.42	0.45
BaAcBcBcB	119	1.17	45.29	1.36	0.41
AaAaAaAcB	108	1.06	46.36	1.28	0.49
BcBcBaAcB	107	1.05	47.41	1.22	0.35
AaAcBcBcB	105	1.03	48.44	1.30	0.39
AaAaAaAaC	95	0.94	49.38	0.93	0.46
BcBaAcBcB	90	0.89	50.27	1.27	0.38
AcBaAaAaA	83	0.82	51.08	1.00	0.47
AcBcBaAaA	81	0.80	51.88	1.12	0.48
AaCdAcBcB	78	0.77	52.65	1.12	0.38
CdAcBcBcB	78	0.77	53.42	1.10	0.36
BaAaAcBcB	68	0.67	54.09	1.20	0.37
AaAaAcBaA	66	0.65	54.74	1.28	0.49
AaAaAcBcB	65	0.64	55.38	1.31	0.45
AaAaAaCdA	65	0.64	56.02	0.95	0.45
BcFdAcBcB	65	0.64	56.66	1.23	0.50
BcBcFdAcB	65	0.64	57.30	1.37	0.65
FdAcBcBcB	65	0.64	57.94	1.13	0.41
AaAaCdAcB	64	0.63	58.57	1.10	0.39
BcBaAaAcB	62	0.61	59.18	1.27	0.51
BcBcBcFdA	62	0.61	59.79	1.39	0.54
AcBcBcBaA	60	0.59	60.38	1.40	0.39
BcBcFdCdA [c]	24			0.58	0.46
BcFdCdAcB	24			0.51	0.23
FdCdAcBcB	28			0.51	0.23
BaAaCdAcB	25			0.69	0.27
AaAaAcFdA	19			0.82	0.32
AaAcFdAcB	17			0.84	0.31
AcFdAcBcB	19			1.25	0.49

[a] Structural codes which contain neither o nor O are listed. The second column lists the number of fragments with the same nine-letter code.

[b] Square-root of mass weighted mean square distance between a pair of fragments having the same structural code is calculated for all pairs. Its average μ and standard deviation σ are given.

[c] Structural codes, not necessarily with a large number of fragments but with sharply defined structures, are given below.

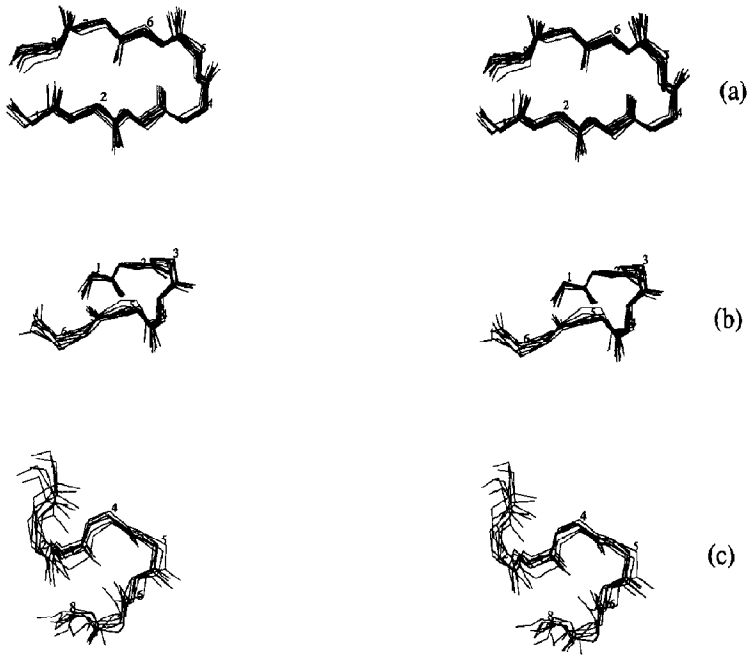

Fig. 7. Stereodrawings of superposed backbone structures of (a) type I′ β-ladder whose structural code is given by BcBcFdCdAcBcB, (b) five-turn whose structural code is BaAaCdAcB, and (c) type II αβ-loop whose structural code is AaAaAcFdAcB.

ARrangement of backbone Fragments) among not only homologous but also non-homologous proteins. This new type of three-dimensional structural similarity appears to be important for understanding the evolution, function and folding mechanism of proteins.

The algorithm to detect common SARFs in a pair of proteins requires to prepare a list of all possible pairs of six-residue fragments with very similar three-dimensional structures. For instance we prepare a list of pairs of fragments with r.m.s. distance smaller than 0.8 Å. When such pairs

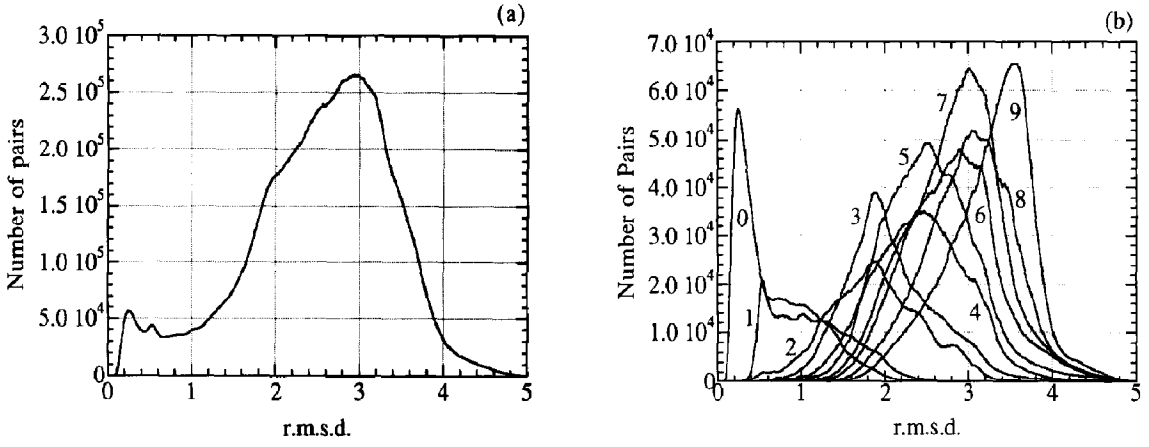

Fig. 8. (a) Distribution of r.m.s. distance (r.m.s.d.) between all possible pairs of six-residue fragments taken irrespective of their structural codes from the 76 proteins. (b) Its breakdown into contributions from pairs of fragments with a given number of different one-letter structural codes.

were calculated for a pair of proteins 4APE and 3HVP [8], 267 pairs were obtained by using 66.0 seconds CPU time on a TITAN 1500 workstation. However, what we learned from Fig. 8 indicates that this search can be shortened by limiting search only for pairs with no or one different one-letter structural code and for pairs with two different one-letter codes involving at least one o or O. According to this procedure we obtained 256 similar pairs in 13.1 seconds CPU time. This means that limitation of the search can be done essentially without loosing an important fraction of pairs.

4. Conclusion

Three-dimensional structures of peptide backbone (ϕ, ψ) and (ψ, ϕ) fragments are found classifiable into five and six clusters, respectively. Protein backbone structural code based on these clusters are found to be useful, for example, to describe various types of four-turns, or to detect local similarities among protein backbone structures.

Acknowledgements

Computation has been done at the Computer Centers of Kyoto University and the Institute for Molecular Science. This work has been supported by grants from the Ministry of Education, Science and Culture of Japan, and from the International Human Frontier Science Program Organization.

References

1 C. Chothia, Nature 357 (1992) 543.
2 W. Kabsch and C. Sander, Biopolymers 22 (1983) 2577.
3 A.W. Burgess, P.K. Ponnuswamy and H.A. Scheraga, Israel J. Chem. 12 (1974) 239.
4 M.H. Lambert and H.A. Scheraga, J. Comp. Chem. 10 (1989) 770.
5 M.J. Dudek and H.A. Scheraga, J. Comp. Chem. 11 (1990) 121.
6 C.M. Wilmot and J.M. Thornton, Protein Eng. 3 (1990) 479.
7 M.J. Rooman, J.P.A. Kocher and S.J. Wodak, J. Mol. Biol. 221 (1991) 961.
8 F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F.J. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, J. Mol. Biol. 112 (1977) 535.
9 A.D. McLachlan, J. Mol. Biol. 128 (1979) 49.
10 IUPAC-IUB, J. Mol. Biol. 52 (1970) 1.
11 D. Kotelchuck and H.A. Scheraga, Proc. Natl. Acad. Sci., U.S.A. 61 (1968) 1163.
12 D. Kotelchuck and H.A. Scheraga, Proc. Natl. Acad. Sci., U.S.A. 62 (1969) 14.
13 P.K. Ponnuswamy and V. Sasisekharan, Biopolymers 10 (1971) 565.
14 P.N. Lewis, F.A. Momany and H.A. Scheraga, Biochim. Biophys. Acta 303 (1973) 211.
15 J.S. Richardson, Adv. Protein Chem. 34 (1981) 1167.
16 B.L. Sibanda and J.M. Thornton, Nature 316 (1985) 170.