Protein Engineering vol.11 no.11 pp.981–990, 1998

Novel method to detect a motif of local structures in different protein conformations

Hiroshi Wako[1] and Takahisa Yamato[2]

School of Social Sciences, Waseda University, Shinjuku-ku, Tokyo 169-8050 and [2]Faculty of Science, Nagoya University, Chikusa-ku, Nagoya 464-8602, Japan

[1]To whom correspondence should be addressed

In order to detect a motif of local structures in different protein conformations, the Delaunay tessellation is applied to protein structures represented by C_α atoms only. By the Delaunay tessellation the interior space of the protein is uniquely divided up into Delaunay tetrahedra whose vertices are the C_α atom positions. Some edges of the tetrahedra are virtual bonds connecting adjacent residues' C_α atoms along the polypeptide chain and others indicate interactions between residues nearest neighbouring in space. The rules are proposed to assign a code, i.e., a string of digits, to each tetrahedron to characterize the local structure constructed by the vertex residues of one relevant tetrahedron and four surrounding it. Many sets comprised of the local structures with the same code are obtained from 293 proteins, each of which has relatively low sequence similarity with the others. The local structures in each set are similar enough to each other to represent a motif. Some of them are parts of secondary or supersecondary structures, and others are irregular, but definite structures. The method proposed here can find motifs of local structures in the Protein Data Bank much more easily and rapidly than other conventional methods, because they are represented by codes. The motifs detected in this method can provide more detailed information about specific interactions between residues in the local structures, because the edges of the Delaunay tetrahedra are regarded to express interactions between residues nearest neighbouring in space.

Keywords: Delaunay tessellation/residue–residue interaction/structural classification/structural codes/structural motifs

Introduction

In order to understand protein structure, a study of commonly occurring local structures in different proteins, which we refer to as a motif, is very important. It is well known that the three-dimensional (3D) structure of proteins has been better conserved during evolution than amino acid sequence. A simple relationship between sequence identity and structural similarity has been found among homologous proteins (Chothia and Lesk, 1986; Sander and Schneider, 1991; Flores *et al.*, 1993; Chelvanayagam *et al.*, 1994). Even if it is hard to show statistically significant sequence similarity, evolutionary relationships between distantly related proteins can be inferred based on 3D structural similarity (Johnson *et al.*, 1990a,b). Several structural alignment methods, which take into account several structural properties besides the amino acid sequence, have been also proposed to improve accuracy of the sequence alignment (Sali and Blundell, 1990; Flores *et al.*, 1993; Holm and Sander, 1993; Chelvanayagam *et al.*, 1994; Russell and Barton, 1994). To go a step further, methods have been developed for systematic structure comparison among not only homologous but also non-homologous proteins. They are suitable for databank searches and clustering (Taylor and Orengo, 1989; Alexandrov *et al.*, 1992; Holm and Sander, 1993, 1996; Alexandrov and Go, 1994; Mizuguchi and Go, 1995; Orengo and Taylor, 1996).

Since most motifs play a functionally or structurally important role, the motif searches may be expected not only to give insight into the relationship between proteins and their possible evolutionary origins, but also to deepen our understanding of the relationships between the amino acid sequence and the 3D structure. Accumulation of many motifs can serve as a database that is helpful for homology modelling, *ab initio* prediction of structure from sequence, and *de novo* design of proteins. Our purpose in this paper is to propose a novel method to detect a motif of local structures in different protein conformations.

It is usually more comprehensive to assume that protein structures are arranged in hierarchical fashion, i.e., amino acid sequence, secondary structure, supersecondary structure, domain and tertiary structure (Schurtz and Schirmer, 1979; Richardson, 1981). Accordingly, it is reasonable to define motifs at each level of this hierarchy. At secondary structure level α-helix, β-strand and various kinds of turns are regarded as motifs. At supersecondary structure level there are various motifs assembling a couple of secondary structure elements such as β-α-β, α-turn-α, parallel- and antiparallel-β-sheet and α-helix bundle. At domain level the motif is a more ambiguous concept. Their classification has not been established yet. In fact, even in FSSP (Holm and Sander, 1994), SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1993, 1996), well-known domain classification databases, their classifications do not coincide with each other for some proteins at present. The motifs in which we are interested are at secondary and supersecondary structure levels, or at a structure level classified in between. As a matter of fact, some of the motifs defined in this paper are parts of secondary and supersecondary structures, and others are structures of similar size not directly related to the secondary structures. We refer to these structure as local structures. The method presented here can provide more detailed information about the motifs at such a level.

A view of protein structures as an assembly of secondary structures is the most popular. Such a picture is considered reasonable, but not trivial. In fact, a module proposed by Go (1985) is defined based on the compactness of a contiguous local region of protein irrespective of the secondary structures. In the method proposed here we do not presuppose the existence of the secondary structures either. We think that the method to detect motifs without the presupposition of secondary structures is significant, even if it may be shown after the analyses that secondary structures are essential elements to describe the protein structure. This attitude is also useful to

analyse the regions with ambiguously defined structures such as N- and C-terminals of secondary structures and irregular (but possibly definite) loop structures.

It is also optional in the analysis of the motifs whether or not they are defined as a contiguous region along the polypeptide chain. In this study we do not give such a restriction to the detection algorithm explicitly. As described below, we will focus our attention on the network of spatially nearest neighbouring residues in protein structures. However, information about sequential connectivity of the amino acid residues is included implicitly through the unique numbering method of vertex residues on the Delaunay tetrahedron proposed below.

In this paper we apply the Delaunay tessellation to protein structures to detect a motif of local structures. Delaunay tessellation has been used for structural analyses of various disordered systems. In most such cases it has served as a valuable tool for structure description (Voloshin *et al.*, 1989; Vaisman *et al.*, 1994). As for protein structures, Yamato *et al.* (1994) employed the Delaunay tessellation to analyse a thermally fluctuating protein structure in molecular dynamics simulation. It was also applied to the analyses of pressure-induced deformations of proteins derived from normal mode analysis (Yamato, 1996; Kobayashi *et al.*, 1997). In these studies the Delaunay tetrahedra are used to define topographical structures and metric of the protein molecule at atomic level.

Recently, Singh *et al.* (1996) made statistical analysis of residue composition of Delaunay tetrahedra and revealed that non-random preferences for certain quadruplets of amino acids to be clustered together. This work was developed further by the same group (Munson and Singh, 1997) to derive empirical multi-body potentials and to apply the obtained potentials to sequence–structure alignment.

The Delaunay tetrahedron is composed of four vertices (C_α atom positions in this study) and six edges connecting them. Since the edges connect essentially the nearest neighbouring C_α atoms in space, the characterization of the networks of the edges can be used to define types of local structures with respect to residue–residue interactions. For such a characterization we propose a method to assign a code, i.e., a string of digits, to each Delaunay tetrahedron, and then show that the local structures with the same code are similar enough to each other to represent a motif.

Materials and methods

Delaunay tessellation

In this work the Delaunay tessellation is applied to the analysis of protein structures. By representing amino acid residues in the protein molecule only by C_α atoms, the protein structure is described as a set of points (C_α atoms) in 3D space. The Delaunay tessellation of the protein structure generates an aggregate of space-filling irregular tetrahedra, or Delaunay simplices (we refer to them as Delaunay tetrahedra or simply as tetrahedra in this paper). To explain the Delaunay tessellation we first describe a related geometric construct, the Voronoi polyhedron. That is, an entire 3D volume is divided into non-overlapping regions called Voronoi polyhedra, each of which is defined as a set of points closest to one particular particle (C_α atom in this case) of the protein. The boundary points of the Voronoi polyhedra are thus equally distant from two particles. Particles whose Voronoi polyhedra share a common boundary are said to be in contact or nearest neighbours of each other. If we connect particles in contact with a line segment, there appear the Delaunay simplices composed of the vertices (i.e., the C_α atoms) and the line segments connecting them. It is well known that the Delaunay simplices in 3D space are always tetrahedra. The complete set of tetrahedra divides up the interior space of the protein into non-overlapping volume elements. It is called Delaunay tessellation. This tessellation uniquely defines all the Delaunay tetrahedra for a given protein structure.

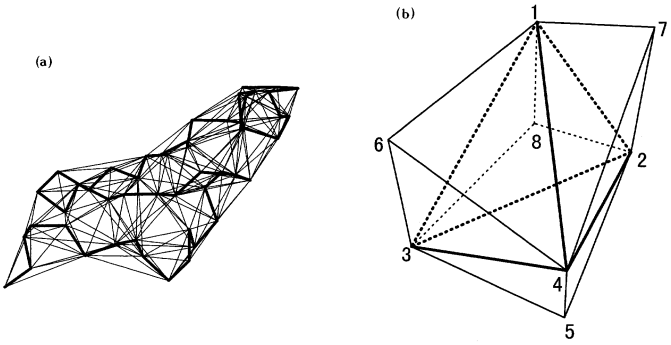

Fig. 1. An example of the Delaunay tessellation of protein and vertex numbering. (**a**) A thick line is a C_α trace of human transforming growth factor α, which consists of 50 amino acid residues (PDB entry ID is 4tgf; Kline *et al.*, 1990). A thin line is an edge of the Delaunay tetrahedron generated by the Delaunay tesselation. (**b**) A central tetrahedron 1234 and four surrounding tetrahedra, 2345, 1346, 1247 and 1238. To the four vertices of the central tetrahedron, digits, 1, 2, 3 and 4, are assigned in increasing order of the corresponding vertex residues' numbers. To the four vertices of the surrounding tetrahedra besides the common vertices with the central one, digits, 5, 6, 7 and 8, are assigned just as shown in this figure. The vertex residues 5, 6, 7 and 8 do not always exist.

In order to calculate the Delaunay tessellation we use the program Qhull (Barber *et al.*, 1995) obtained by anonymous ftp (geom.umn.edu/pub/software). In the calculated tessellation for a given protein structure, however, we found that some pairs of C_α atoms too distant from each other are connected in some geometrically irregular regions, such as on the surface, active site crevasses and N- and C- terminals. Since we assume that the edges of the Delaunay tetrahedra connecting two amino acid residues reflect some interactions between these residues in this work, we took into account only the Delaunay tetrahedra in which all of the four edges are shorter than a given cut-off distance (10 Å in this study). An example of the Delaunay tessellation of a protein molecule is shown in Figure 1a.

The Delaunay tessellation is just a geometrical operation without any explicit consideration of properties specific to protein structures. It should be emphasized that the chain connectivity and secondary structures are not taken into computation of the tessellation, even though they are taken into account implicitly, because the residues in close contact are very likely to be connected as the edges of the Delaunay tetrahedra.

Code assignment to tetrahedron

Consider a Delaunay tetrahedron T_0. The amino acid residue numbers at the four vertices of the tetrahedron are denoted $v_1(T_0)$, $v_2(T_0)$, $v_3(T_0)$ and $v_4(T_0)$. Here we require the following rule for putting the suffixes 1 to 4, i.e., $v_1(T_0) < v_2(T_0) < v_3(T_0) < v_4(T_0)$. Accordingly, we can uniquely number four vertices of any Delaunay tetrahedron with the digits 1 to 4.

We also consider the surrounding tetrahedra which share one of the facets (triangles) of T_0. At most it is possible for four tetrahedra to exist, T_5, T_6, T_7 and T_8, although they do

not always do so. If they exist, we denote the residue numbers at the vertices of these tetrahedra not contained by T_0, $v_5(T_0)$, $v_6(T_0)$, $v_7(T_0)$ and $v_8(T_0)$, respectively. The four tetrahedra, 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. It should be noticed that the suffixes 5 to 8 are also uniquely assigned to the vertex residues in relation to suffixes 1 to 4 (see Figure 1b). It should be also noticed that it frequently occurs that some of the vertices v_5 to v_8 share the same residue.

Then we assign a code, a string of digits, $c(T_0)$, to the tetrahedron T_0 according to the following rules:

(1) Arrange v_1 to v_8 in increasing order.

(1a) If $v_a = v_b$ and $a < b$ for a, $b \geqslant 5$ (i.e., some vertex residues of T_a and T_b are coincident with each other), arrange v_a, v_b in this order.

(1b) If v_a for $a \geqslant 5$ does not exist (i.e., tetrahedron T_a does not exist), v_a is not included in the arrangement.

(2) If v_1 to v_8 are arranged as v_a, v_b, v_c, v_d, v_e, v_f, v_g, v_h (the digits 1 to 8 may be assigned to suffixes a to h in various permutation), code 'abcdefgh' is assigned to this tetrahedron. If the number of the vertices is less than 8, the number of digits in the code is also less than 8 according to the above rule (1b). Since the digits 1 to 4 of the suffixes always appear in this order, the code has at least four digits.

Let us show some examples. If $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\} = \{32, 34, 35, 36, 45, 59, 23, 22\}$, then the code is 87123456 ($v_8 < v_7 < v_1 < v_2 < v_3 < v_4 < v_5 < v_6$). For {13, 14, 16, 17, 15, 82, 82, 15}, the code is 12583467 ($v_1 < v_2 < v_5 = v_8 < v_3 < v_4 < v_6 = v_7$). This is the case where $v_5 = v_8$ and $v_6 = v_7$. For {51, 60, 62, 63,, 12,, 11}, the code is 861234 ($v_8 < v_6 < v_1 < v_2 < v_3 < v_4$). This is the case where v_5 and v_7 are missing. Some other examples can be found in Results.

There are 1680 possible codes of 8 digits. There are also 840, 180, 20 and 1 possible codes of 7, 6, 5 and 4 digits, respectively. The total number of possible codes is 2721.

As shown in Results, however, this code $c(T_0)$ is not enough to detect a motif of local structures. We found that it is better to take into account the tetrahedra T_5, T_6, T_7 and T_8, neighbouring T_0. The codes are assigned to these tetrahedra in the same manner as T_0. Including these four codes we assign a set of the codes $c(T_0)$: $c(T_5)$: $c(T_6)$: $c(T_7)$: $c(T_8)$ to the tetrahedron T_0. We refer to the former code as a single tetrahedron (ST) code, and to the latter as a nearest neighbour tetrahedra (NNT) code, in this paper.

In the NNT code, at most 20 residues are taken into account, since each of the four tetrahedra, T_5, T_6, T_7 and T_8, has three nearest neighbouring tetrahedra, T_9 to T_{11}, T_{12} to T_{14}, T_{15} to T_{17} and T_{18} to T_{20}, respectively, except for T_0. The residue numbers at the vertices of T_9 to T_{20} not included in T_5 to T_8 are referred to as v_9 to v_{20}, respectively. Since some of the residues v_9 to v_{20} do not exist, or share the same vertices similarly to v_5 to v_8, the net number of residues related to the NNT code is not necessarily equal to 20, but usually less than 20. This is the size of the local structures considered in this paper.

(The vertex residues for the tetrahedra T_9 to T_{20} are explicitly given in the followings: $T_9 = \{v_3, v_4, v_5, v_9\}$, $T_{10} = \{v_2, v_4, v_5, v_{10}\}$, $T_{11} = \{v_2, v_3, v_5, v_{11}\}$, $T_{12} = \{v_3, v_4, v_6, v_{12}\}$, $T_{13} = \{v_1, v_4, v_6, v_{13}\}$, $T_{14} = \{v_1, v_3, v_6, v_{14}\}$, $T_{15} = \{v_2, v_4, v_7, v_{15}\}$, $T_{16} = \{v_1, v_4, v_7, v_{16}\}$, $T_{17} = \{v_1, v_2, v_7, v_{17}\}$, $T_{18} = \{v_2, v_3, v_8, v_{18}\}$, $T_{19} = \{v_1, v_3, v_8, v_{19}\}$, $T_{20} = \{v_1, v_2, v_8, v_{20}\}$).

Table I. The most abundant ST codes and their residue number patterns

Code	Number of tetrahedra	Residue number pattern							
		v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
68123457	18 094 (total)								
	14 493	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	573	i	i+1	j	k	k+1	i–1	k+1	i–1
	482	i	i+1	i+2	i+3	i+4	i–1	j	i–1
	475	i	i+1	i+2	j	k	l	m	n
	296	i	i+1	i+2	i+3	i+4	j	i+4	i–1
	1775	other patterns							
67125834	5584 (total)								
	1586	i	i+1	i+3	i+4	i+2	j	j	i+2
	1272	i	i+1	j	j+1	j–1	i–1	i–1	j–1
	1178	i	i+1	i+3	i+4	i+2	j	k	i+2
	1548	other patterns							
12583467	4762 (total)								
	924	i	i+1	j	j+1	j–1	k	m	j–1
	894	i	i+1	j	j+1	j–1	k	k	j–1
	847	i	i+1	j	j+1	k	j+2	m	i+2
	561	i	i+1	j	j+1	j–1	j+2	j+2	j–1
	419	i	i+1	i+3	i+4	i+2	i+5	i+5	i+2
	1117	other patterns							
125834	4614 (total)								
	3683	i	i+1	i+3	i+4	i+2	–	–	i+2
	931	other patterns							

Proteins analyzed

The proteins analysed in this paper are given in Appendix 1.

Results

By the Delaunay tessellation 208 434 tetrahedra are obtained from 293 proteins. The two kinds of codes, ST and NNT, are assigned to each tetrahedron according to the rules described above. Of the 2721 possible ST codes, 405 are not found, 206 are assigned only to one tetrahedron and 131 to two tetrahedra.

On the contrary, the most abundant ST code is 68123457, which is assigned to 18 094 tetrahedra. The next most abundant ST codes are 67125834, 12583467, 125834, 1258346 and 12683457, which are assigned to 5584, 4762, 4614, 1782 and 1771 tetrahedra, respectively.

In Table I the residue number patterns at the eight vertices v_1 to v_8, $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$, denoted by $\{v_1$–$v_8\}$ hereinafter, are shown for the top four most abundant ST codes. The residue number patterns are used as implication of 3D structures at the first glance (the structures with the same residue number patterns are similar to each other in general; however, it is necessary to make certain of the structural similarity by their superposition). Residue number v_1 is set to i, and then the relative residue numbers are shown for v_2 to v_8. If the residue number is regarded to have no relation with v_1, different characters j, k, ... are used.

Table I shows that there are various residue number patterns for one code. As shown below, the first code 68123457 corresponds essentially to part of the α-helix structure. However, the second and third codes, 67125834 and 12583467, correspond to either part of the α-helix or β-sheet. This means that different local structures are assigned to the common ST code. Consequently, the ST code alone is not enough to distinguish the local structures.

Table II. The most abundant NNT codes and their residue number patterns

Code	Number of tetrahedra	Residue number pattern							
		v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
125834:68123457:0:0:68123457									
	3355 (total)								
	3271	i	i+1	i+3	i+4	i+2	–	–	i+2
	78	i	i+1	j	j+2	k	–	–	j+1
	6	other patterns							
68123457:68123457:125834:125834:68123457									
	905 (total)								
	902	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	3	i	i+1	i+2	i+3	i+4	i–3	i+4	i–3
68123457:68123457:1258346:7125834:68123457									
	574 (total)								
	574	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
t68123457:68123457:125834:1258347:68123457									
	536 (total)								
	536	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
68123457:68123457:125834:67125834:68123457									
	378 (total)								
	375	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	3	other patterns							
68123457:68123457:125834:125834:68123457									
	360 (total)								
	357	i	i+1	i+2	i+3	i+4	i–1	i+4	i–1
	3	other patterns							
123457:68123457:0:67125834:0									
	355 (total)								
	341	i	i+1	i+2	i+3	i+4	–	i+4	–
	14	other patterns							

In Table II the residue number patterns $\{v_1–v_8\}$ are given for the top seven most abundant NNT codes. Although each NNT code corresponds to only one residue number pattern in general, some codes correspond to more than one pattern. Even in such a case, however, the exceptional patterns are only small variations of the major one. Completely different patterns are very few. (The residue number patterns for the second to seventh codes shown in Table II are incidentally identical. There are various residue number patterns for other codes not shown here). Residue numbers v_9 to v_{20} are not shown in Table II, because it is too lengthy to show them and because they have more varieties of patterns in some cases.

In the following we demonstrate with some examples that motifs of local structures are well detected with the NNT codes. In the analyses we consulted the database SCOP (Murzin *et al.*, 1995) and examined the results with the help of the program, PROMOTIF (Hutchinson and Thornton, 1996).

α-Helix

The top 20 most abundant ST codes can be classified into two groups except for four codes. In one group the codes include a sequence 125834; for example, 67125834 (5584), 12583467 (4762), 125834 (4614), 1258346 (1782), 7125834 (1612), 61258347 (1532), 6125834 (1416), 1258347 (1413) etc. (the numbers of tetrahedra found with these codes, N_{ST}, are shown in the parentheses). In another group the codes include a sequence 1234: for example, 68123457 (18 094), 681234 (1458), 123457 (1377), 6812345 (1227), 8123457 (1045) etc. The four exceptional codes are 12683457 (1771), 68125734 (1427), 68123574 (1131) and 16823457 (989), which are discussed in the β-sheet subsection below.

The typical residue number pattern $\{v_1–v_8\}$ for the codes with 125834 is {i, i+1, i+3, i+4, i+2, j, k, i+2}. The pattern

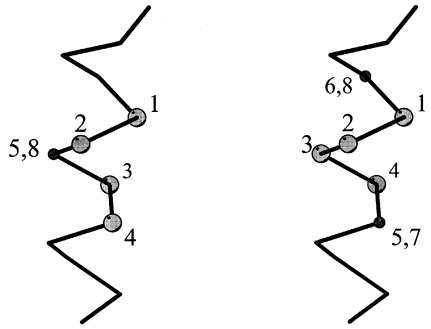

Fig. 2. Vertex residues on α-helix. (**a**) Code 125834; (**b**) code 68123457. Figures 2–8 were drawn by MOLSCRIPT (Kraulis, 1991).

for the code 68123457 is {i, i+1, i+2, i+3, i+4, i–1, i+4, i–1}. In dual expressions, the vertices v_1, v_2, $v_5 = v_8$, v_3 and v_4 correspond to the residue numbers, i to i+4, respectively, for the former code (symbol = is used to indicate that the two vertex residues, v_5 and v_8 in this case, are identical). v_6 and v_7 do not have definite residue numbers relative to v_1. For the latter code, the vertices $v_6 = v_8$, v_1, v_2, v_3, v_4, and $v_5 = v_7$ correspond to the residue numbers, i–1 to i+4, respectively. The local structures corresponding to both the codes are one to two turns of α-helix, as shown in Figure 2. In other words, the two abundant ST codes, 68123457 and 125834, are principal codes for the α-helix.

The most abundant NNT code,

NNT code = 125834: 68123457: 0: 0: 68123457
$\{v_1–v_8\}$ = {i, i+1, i+3, i+4, i+2,,, i+2}

contains the above two ST codes (see Table II). It represents part of the α-helix structure composed of tetrahedra T_0, T_5 and T_8. Tetrahedra T_6 and T_7 are missing. The residue numbers constructing T_0, T_5 and T_8 are (i, i+1, i+3, i+4), (i+1, i+2, i+3, i+4) and (i–1, i, i+1, i+2), respectively. $c(T_6) = c(T_7) = 0$ means that there are no residues near this part of the α-helix.

For comparison we give an example that has a code with the same $c(T_0)$, $c(T_5)$ and $c(T_8)$ as the above NNT code, but with nonzero $c(T_6)$ and $c(T_7)$ [$c(T_0)$ = 12583467 rather than 125834, because T_6 and T_7 exist]. The NNT code, the residue number pattern $\{v_1–v_8\}$, the number of tetrahedra with this NNT code found in the set of 293 proteins, N_{NNT}, and some examples are shown in Table III(a). Fifteen out of 40 structures, which are superposed with respect to the vertex residues v_1 to v_8, are shown in Figure 3. While residues $v_1–v_5$ and v_8 are on the same α-helix, $v_6 = v_7$ are remote from them along the chain ($v_6 = v_7 > v_1$). The residues v_9 to v_{20} and those neighbouring v_1 to v_{20} along the polypeptide chain are also included in Figure 3. Although all the residues are not well fitted to each other, because the superposition is performed with respect only to v_1 to v_8, these local structures seem to have a common feature with each other to represent a motif. This is supported by the fact that favorable amino acid types are limited at some vertices. For example, hydrophobic amino acid residues are favorable at $v_6 = v_7$ (out of 40 structures with this code, 13, 4, 4, 2, 3 and 2 are Ala, Val, Leu, Ile, Gly and Pro, respectively) and hydrophilic amino acid residues or those with a small side chain are favorable at v_2 (11, 5, 4, 2 and 3 are Glu, Asp, Lys, Arg and Asn, and 5 and 3 are Ala and Gly, respectively).

The residue number patterns of the vertex residues $v_6 = v_7$, v_{12} and v_{14} in the segments located on the right of the α-

Table III. Examples of motifs: NNT codes, residue number patterns and vertex residues[a]

Protein[b]	Vertex residue[c]																			
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	v_{15}	v_{16}	v_{17}	v_{18}	v_{19}	v_{20}
(a) 12583467: 68123457: 16823745: 126834: 68123457, {v_1–v_8} = {i, i+1, i+3, i+4, i+2, j, j, i+2}, N_{NNT} = 40																				
3chy	66 L	67 E	69 L	70 K	68 L	101 A	101 A	68 L	71 T	71 T	66 L	102 G	67 E	98 A	–	69 L	–	70 K	65 G	65 G
1gox	235 A	236 E	238 A	239 R	237 D	275 A	275 A	237 D	240 L	240 L	235 A	276 A	236 E	272 V	–	238 A	–	239 R	234 T	234 T
1ads	28 T	29 E	31 V	32 K	30 A	57 A	57 A	30 A	33 V	33 V	28 T	58 I	29 E	54 V	–	31 V	–	32 K	27 V	27 V
5p21	69 D	70 Q	72 M	73 R	71 Y	103 V	103 V	71 Y	74 T	74 T	69 D	104 K	70 Q	100 I	–	72 M	–	73 R	68 R	68 R
1wsyA	82 F	83 E	85 L	86 A	84 M	121 V	121 V	84 M	87 L	87 L	82 F	122 G	83 E	118 C	–	85 L	–	86 A	81 C	81 C
(b) 12576834: 67125834: 16823457: 68123574: 12583467, {v_1–v_8} = {i, i+1, j+1, j+2, i+2, j, i+2, j}, N_{NNT} = 24																				
2mnr	266 L	267 A	294 M	295 S	268 M	293 P	268 M	293 P	269 P	266 L	269 P	314 H	314 H	267 A	294 M	244 Q	244 Q	292 I	295 S	292 I
5timA	124 I	125 A	164 I	165 A	126 C	163 V	126 C	163 V	127 I	124 I	127 I	209 L	209 L	125 A	164 I	92 V	92 V	162 V	165 A	162 V
4enl	318 V	319 A	342 L	343 L	320 D	341 A	320 D	341 A	323 T	318 V	323 T	370 M	370 M	319 A	342 L	295 E	295 E	339 A	343 L	339 A
1chrA	243 M	244 A	266 F	267 S	245 D	265 V	245 D	265 V	248 L	243 M	248 L	294 Y	294 Y	244 A	266 F	220 E	220 E	263 V	267 S	263 V
1btc	338 I	339 L	377 V	378 A	340 N	376 R	340 N	376 R	341 F	338 I	341 F	414 F	414 F	339 L	377 V	293 A	293 A	375 I	378 A	375 I
(c) 12357468: 67125834: 1782354: 12576834: 12368475, {v_1–v_8} = {i, i+1, i+2, j, j–1, k, j–1, k}, N_{NNT} = 6																				
5p21	5 K	6 L	7 V	56 L	55 I	75 G	55 I	75 G	8 V	5 K	8 V	71 Y	–	6 L	7 V	54 D	54 D	77 G	56 L	76 E
2cmd	2 K	3 V	4 A	31 S	30 L	70 A	30 L	70 A	5 V	2 K	5 V	66 A	–	3 V	4 A	29 E	29 E	72 V	31 S	71 D
1s01	27 K	28 V	29 A	91 Y	90 L	119 M	90 L	119 M	30 V	27 K	30 V	114 A	–	28 V	29 A	89 S	89 S	121 V	91 Y	120 D
3chy	7 K	8 F	9 L	34 E	33 V	51 Y	33 V	51 Y	10 V	7 K	10 V	45 K	–	8 F	9 L	32 N	32 N	53 F	34 E	52 G
1glt	3 K	4 L	5 G	36 H	35 L	80 L	35 L	80 L	6 I	3 K	6 I	77 L	–	4 L	5 G	34 E	34 E	82 V	36 H	81 D
(d) 81236457: 68172345: 68172345: 12583467: 12368475, {v_1–v_8} = {i, i+1, j, j+2, j+3, j+1, j+3, i–1}, N_{NNT} = 6																				
2reb	90 C	91 A	138 V	140 V	141 I	139 D	141 I	89 T	187 T	90 C	132 L	187 T	88 K	89 T	138 V	142 V	142 V	115 L	139 D	115 L
1glt	4 L	5 G	80 L	82 V	83 I	81 D	83 I	3 K	110 T	4 L	77 L	110 T	2 I	3 K	80 L	84 L	84 L	36 H	81 D	36 H
8atcA	157 V	158 A	222 V	224 I	225 L	223 D	225 L	156 H	261 M	157 V	219 M	261 M	155 L	156 H	222 V	226 Y	226 Y	185 Y	223 D	185 Y
1sbp	7 N	8 V	58 A	60 T	61 V	59 D	61 V	6 L	226 V	7 N	52 V	226 V	5 L	6 L	58 A	62 T	62 T	41 S	59 D	41 S
2cmd	3 V	4 A	70 A	72 V	73 V	71 D	73 V	2 K	112 A	3 V	67 L	112 A	1 M	2 K	70 A	74 L	74 L	31 S	71 D	31 S
(e) 68123457: 68123457: 61258347: 125834: 68123457, {v_1–v_8} = {i, i+1, i+2, i+3, i+4, i–1, i+4, i–1}, N_{NNT} = 41																				
4fxn	97 R	98 D	99 F	100 E	101 E	96 M	101 E	96 M	102 R	97 R	102 R	84 L	114 P	98 D	99 F	–	–	95 W	100 E	95 W
2cyp	89 F	90 K	91 F	92 L	93 E	88 G	93 E	88 G	94 P	89 F	94 P	46 L	108 F	90 K	91 F	–	–	87 N	92 L	87 N
4ts1A	99 A	100 R	101 I	102 K	103 E	98 S	103 E	98 S	104 Q	99 A	104 Q	69 V	122 N	100 R	101 I	–	–	97 W	102 K	97 W
2ctc	224 K	225 S	226 A	227 V	228 E	223 A	228 E	223 A	229 A	224 K	229 A	202 L	240 Y	225 S	226 A	–	–	222 V	227 V	222 V
3rubS	84 L	85 A	86 E	87 V	88 E	83 V	88 E	83 V	89 E	84 L	89 E	42 L	101 I	85 A	86 E	–	–	82 Q	87 V	82 Q
(f) 81623574: 67125834: 81276345: 16825734: 81276345, {v_1–v_8} = {i, j, j+1, j+4, j+3, i+3, j+3, i–1}, N_{NNT} = 9																				
3grs	35 S	346 G	347 R	350 A	349 L	38 R	349 L	34 A	348 K	35 S	348 K	351 H	39 A	34 A	347 R	153 L	153 L	343 I	38 R	31 G
1pda	234 M	287 G	288 I	291 A	290 L	237 R	290 L	233 A	289 S	234 M	289 S	292 E	238 L	233 A	288 I	261 A	261 A	284 E	237 R	230 A
1mioC	58 C	185 G	186 H	189 A	188 I	62 V	188 I	57 G	187 H	58 C	187 H	190 N	63 M	57 G	186 H	174 C	174 C	183 S	62 V	55 Y
1pfkA	19 A	264 A	265 S	268 G	267 M	22 G	267 M	18 A	266 R	19 A	266 R	269 A	23 V	18 A	265 S	122 L	122 L	261 R	22 G	15 G
2dri	19 L	237 P	238 D	241 G	240 I	22 G	240 I	18 S	239 Q	19 L	239 Q	242 A	23 A	18 S	238 D	88 L	89 D	236 L	22 G	15 F
(g) 681234: 0: 71258346: 0: 123457, {v_1–v_8} = {i, i+1, i+2, i+3,, i–1,, i–1}, N_{NNT} = 18																				
5fbpA	188 P	189 A	190 I	191 G	–	187 D	–	187 D	–	–	–	192 E	186 L	189 A	–	–	–	–	191 G	–
2gb1	47 D	48 A	49 T	50 K	–	46 D	–	46 D	–	–	–	51 T	45 Y	48 A	–	–	–	–	50 K	–
1aapA	25 V	26 T	27 E	28 G	–	24 D	–	24 D	–	–	–	29 K	23 F	26 T	–	–	–	–	28 G	–
8catA	171 P	172 Q	173 T	174 H	–	170 N	–	170 N	–	–	–	175 L	169 R	172 Q	–	–	–	–	174 H	–
1cseI	58 P	59 G	60 T	61 N	–	57 N	–	57 N	–	–	–	62 V	56 Y	59 G	–	–	–	–	61 N	–
(h) 571234: 61857234: 0: 18723564: 0, {v_1–v_8} = {i, i+1, i+2, i+3, j,, j, }, N_{NNT} = 10																				
2cp4	313 K	314 K	315 G	316 D	301 L	–	301 L	–	300 I	313 K	302 T	–	–	–	–	315 G	312 L	304 D	–	–
2hpdA	348 E	349 K	350 G	351 D	335 A	–	335 A	–	334 Y	348 E	336 K	–	–	–	–	350 G	347 L	338 D	–	–
1glaF	114 K	115 V	116 G	117 D	61 A	–	61 A	–	60 V	114 K	62 P	–	–	–	–	116 G	113 V	64 D	–	–
1hdxA	84 K	85 P	86 G	87 D	73 V	–	73 V	–	72 I	84 K	74 E	–	–	–	–	86 G	83 V	76 V	–	–
1plc	65 A	66 K	67 G	68 E	31 N	–	31 N	–	30 K	65 A	32 N	–	–	–	–	67 G	63 L	35 F	–	–
(i) 8123574: 67125834: 0: 86125734: 6182345, {v_1–v_8} = {i, i+1, i+2, j, j–1,, j–1, k}, N_{NNT} = 7																				
2sn3	40 Y	41 C	42 Y	47 W	46 C	–	46 C	20 G	45 A	40 Y	45 A	–	–	–	42 Y	39 G	25 C	19 L	–	21 E
1ixa	61 S	62 C	63 K	72 W	71 C	–	71 C	51 C	70 E	61 S	70 E	–	–	–	63 K	60 G	56 C	50 Q	–	54 N
1pnh	18 L	19 G	20 K	29 V	28 C	–	28 C	9 Q	27 E	18 L	27 E	–	–	–	20 K	17 L	12 C	8 C	–	13 R
2tgi	77 C	78 C	79 V	112 S	111 C	–	111 C	45 A	110 K	77 C	110 K	–	–	–	79 V	76 P	48 C	44 C	–	46 G
1hcc	46 K	47 C	48 L	53 S	52 W	–	52 W	26 G	51 K	46 K	51 K	–	–	–	48 L	45 A	8 P	25 Y	–	27 E
1avdA	82 Q	83 C	84 F	94 K	93 L	–	93 L	61 Q	92 V	82 Q	92 V	–	–	–	84 F	81 G	64 F	4 C	–	62 P
1ppn	109 G	110 V	111 R	210 V	209 P	–	209 P	73 Q	208 Y	109 G	208 Y	–	–	–	111 R	107 T	72 L	69 W	–	76 A

[a]Only five examples are given for each motif except for (i).
[b]Protein names are given by PDB entry codes.
[c]Underlined residues are discussed in the text.

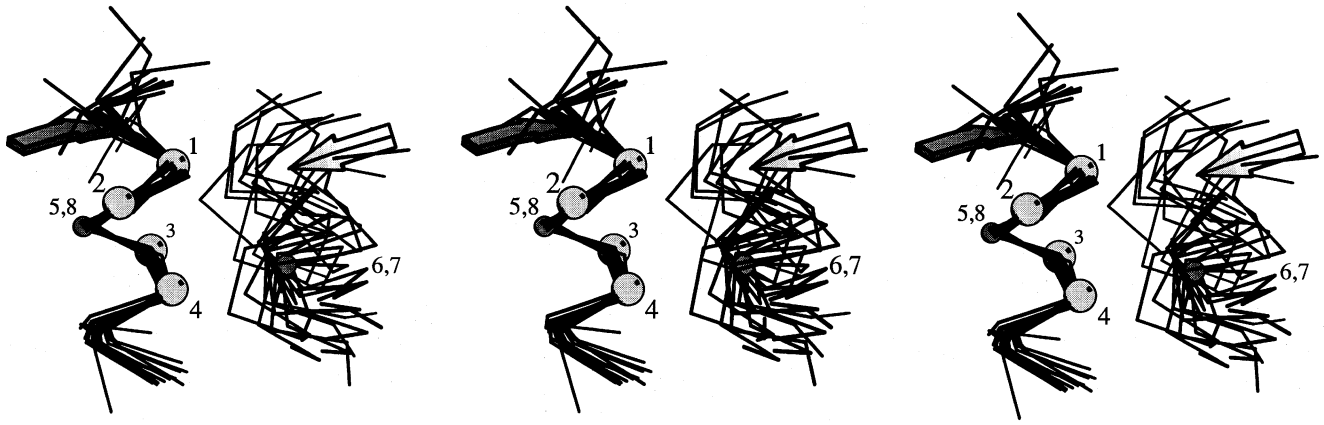

Fig. 3. Stereoscopic view of the local structures with NNT code 12583467: 68123457: 16823745: 126834: 68123457. Fifteen structures with this code are superposed with respect to v_1 to v_8. The residues corresponding to v_1 to v_{20} [given in Table III (a)] and additional residues neighbouring them along the polypeptide chain are included in the structures drawn here. The four large balls indicate the residues v_1 to v_4. The small balls indicate the residues v_5 to v_8. In this case, however, v_5 and v_8, and v_6 and v_7 are the same residues, respectively. The balls indicate the locations of the corresponding residues for the first protein in the list of Table III (a). The arrows do not represent β-sheet, but indicate the direction of the polypeptide chains.

helices in Figure 3 are $(v_6 = v_7, v_{12}, v_{14}) = (j, j+1, j–3)$. This pattern implies that the structures of the segments are also α-helix. In actual fact, they are part of the α-helix. However, some are C-terminal and others are in the middle part of the α-helix. In addition, their relative spatial positions to the α-helices on the left differ (remember that the superposition of the structures was performed with respect only to the vertex residues v_1 to v_8). Nonetheless, the fact that favorable amino acid types are limited at $v_6 = v_7$, as described above, indicates an importance of some particular interactions of the residues at $v_6 = v_7$ in the segment on the right with the residues in the α-helix on the left in Figure 3.

The second to seventh NNT codes in Table II commonly include the ST code 68123457 for tetrahedra T_0, T_5 and T_8, while either $c(T_6)$ or $c(T_7)$ is slightly different from each other. The varieties in $c(T_6)$ and $c(T_7)$ reflect the differences in the environment of the α-helices. This fact indicates that the α-helices can be classified according to the NNT codes that reflect the residues surrounding them.

The codes related to N- and C-terminal structures of the α-helix are also found. An example for the C-terminal one is:

NNT code = 12583467: 681234: 57168234: 12683457: 68123457,
$\{v_1–v_8\} = \{i, i+1, i+3, i+4, i+2, i+5, i+5, i+2\}$.

(data is not shown here). $N_{NNT} = 26$. Residues v_1, v_2, v_3 and v_5 take part in forming α-helix, and residue v_4 terminates it. v_4 is a special residue. Out of 26, 21 are Gly, three are Asn, and the remaining are Asp and His. These are a typical example of Gly-based motifs that cap the C-terminal end of α-helices, so-called C-cap of α-helices (Harper and Rose, 1993; Aurora *et al.*, 1994).

Residues v_{12} and v_{14} are on another segment, similar to Figure 3. The structures of these segments are much more different from each other than those in Figure 3, while the structures of C-terminal regions, where residues v_1, v_2, v_3 and v_5 are located, are well fitted to each other. The residue number patterns of the segments have also more varieties; $(v_{12}, v_{14}) = (i, i), (i, i+2), (i, i+3), (i, i+4), (i, i+5)$ etc.

β-Sheet

There are many codes to represent parallel and antiparallel β-sheet. Similar to the α-helix, these codes correspond to parts of the β-sheets. The most abundant ST codes related to the β-sheet are 67125834, 12583467, 12683457, 68125734, 68123574 and 12576834 (N_{ST}= 5584, 4762, 1771, 1427, 1131 and 813, respectively). Some of them have already appeared above in the top 20 most abundant ST codes. As shown in Table I, the first two codes correspond to part of the α-helix, too. The residue number patterns for these six codes that correspond to the β-sheet (other patterns that do not correspond to the β-sheet for these codes are omitted here) are $\{v_1–v_8\} = \{i, i+1, j, j+1, j–1, i–1, i–1, j–1\}$, $\{i, i+1, j, j+1, i+2, j+2, j+2, i+2\}$, $\{i, i+1, j, j+1, j+2, j–1, j+2, j–1\}$, $\{i, i+1, j, j+1, i+2, i–1, i+2, i–1\}$, $\{i, i+1, i+2, j, j–1, k, j–1, k\}$, and $\{i, i+1, j, j+1, i+2, j–1, i+2, j–1\}$, respectively. Here, the residues, i and j, are on different β-strands.

An example for parallel β-sheet is given in Figure 4 and Table III (b). The residues on the two parallel β-strands, $(i, i+1, i+2)$ and $(j, j+1, j+2)$, are $(v_1, v_2, v_5 = v_7)$ and $(v_6 = v_8, v_3, v_4)$, respectively. There are four strands in Figure 4. Residues v_1 to v_8 are on the two central strands, while residues $v_{12} = v_{13}$ and $v_{16} = v_{17}$ are on the left and right strands, respectively. The folding types of proteins are mainly TIM barrels. α/β and all-β proteins are also included.

For local structures related to the parallel β-sheet, we found the cases where limited amino acid types are favoured by some particular vertices. Let us show two examples. [The structures are not shown here. Only the residues corresponding to v_1 to v_{20} are shown in Table III (c) and (d)]. In Table III (c) five of six v_1's are Lys, and the remaining one is Thr. The residues for v_2 and v_5 are hydrophobic; out of six, two, two, one and one are Leu, Val, Ile and Phe for v_2, and four, one and one are Leu, Val and Ile for v_5, respectively. In Table III (d) five of six v_6's are Asp, and the remaining one is Gly. Incidentally the folding types of all the proteins containing these motifs are α/β.

Next we give an example for anti-parallel β-sheet:

NNT code = 67125834: 6812345: 68125734: 8123574: 12683457,
$\{v_1–v_8\} = \{i, i+1, j, j+1, j–1, i–1, i–1, j–1\}$

(the data is not shown here). $N_{NNT} = 39$. The residues on the two anti-parallel β-strands, $(i–1, i, i+1)$ and $(j+1, j, j–1)$ are $(v_5 = v_6, v_1, v_2)$ and $(v_4, v_3, v_5 = v_8)$, respectively. This motif consists of four strands. Residues v_1 to v_8 are on the two central strands, while residues v_{17} and v_9 are on the outside

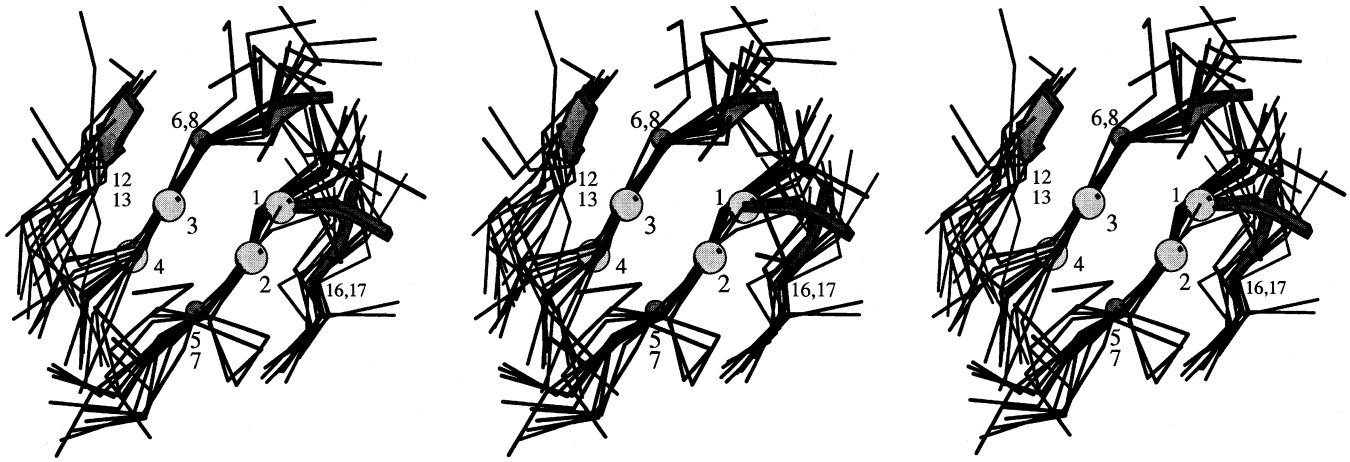

Fig. 4. Stereoscopic view of the local structures with NNT code 12576834: 67125834: 16823457: 68123574: 12583467. The amino acid residues are shown in Table III (b). Fourteen structures are superposed. See also the caption of Figure 3.

Fig. 5. Stereoscopic view of the local structures with NNT code 68123457: 68123457: 61258347: 125834: 68123457. The amino acid residues are shown in Table III (e). Thirteen structures are superposed. See also the caption of Figure 3.

strands. As for folding classes of the proteins, α+β, all-β and multi-domain are dominant.

Supersecondary structures

By the term supersecondary structure we mean a local structure composed of more than one secondary structure element, α-helix and/or β-strand in this paper. Three examples are given here. (Although the β-sheets are a supersecondary structure in this sense, they have been shown above according to the conventional classification).

The β-α-β is the most typical supersecondary structure (Richardson, 1981). Figure 5 and Table III (e) show an example related to this motif. 68123457 is a principal code of α-helix as described above. The two groups of β-strands lay under the α-helix in Figure 5. The vertex residues v_{12} and v_{13} are on the left and right groups of the β-strands, respectively. The relatively wide distribution of the β-strands reflect the fact that their relative spatial positions to the α-helices are not determined strictly in the β-α-β motif. As a matter of course, folding classes of most proteins are α/β.

Two anti-parallel α-helices is found with the following code:

NNT code = 7128354: 71823546: 0: 1682354: 1283457,
$\{v_1–v_8\}$ = {i, i+1, j, j+4, j+1,, i–3, i+4},

N_{NNT} = 12. There is one exceptional case of $\{v_1–v_8\}$ = {i, i+1, j, j+4, j+1,, i–4, i+4}. The folding types of the whole proteins are all-α.

In Figure 6 and Table III (f) an example related to two parallel α-helices over a β-strand is shown. There is one exceptional case of $\{v_1–v_8\}$ = {i, j, j+1, j+4, j+3, i+4, j+3, i–1} for 1mioC. A one-residue insertion caused this difference. It is observed in the lower part of the right α-helix in Figure 6. While most of the vertex residues v_1 to v_{20} are on either of the two α-helices, residues $v_{16} = v_{17}$ are on the β-strand under the α-helices in Figure 6. It is remarkable that the amino acid type of v_2 is Gly for seven of nine, and those of the remaining two are Ala and Pro. The amino acid residues with a small side chain are favorable for v_4 and v_8. The amino acid type of v_4 is Ala for five of nine and those of the remaining four are Gly, Val and Ile. The amino acid type of v_8 is Ala and Gly for four and three of nine, respectively. Those of the remaining two are Ser and Thr. The folding classes of the proteins are either α/β or α+β.

Miscellaneous motifs

There are many interesting motifs not related to secondary structures, each of which is a set of local structures with the same code. Only four examples are shown here.

Table III (g) shows an example related to a turn (the structure is not shown here). Residues v_1 to v_4 form a β-turn of type I (type IV in few cases). The β-strands preceding and succeeding this turn form an antiparallel β-sheet with each other. The residues Asp and Asn are favorable at $v_6 = v_8$.

Figure 7 and Table III (h) show an example related to a

Fig. 6. Stereoscopic view of the local structures with NNT code 81623574: 67125834: 81276345: 16825734: 81276345. The amino acid residues are shown in Table III (f). Nine structures are superposed. See also the caption of Figure 3.

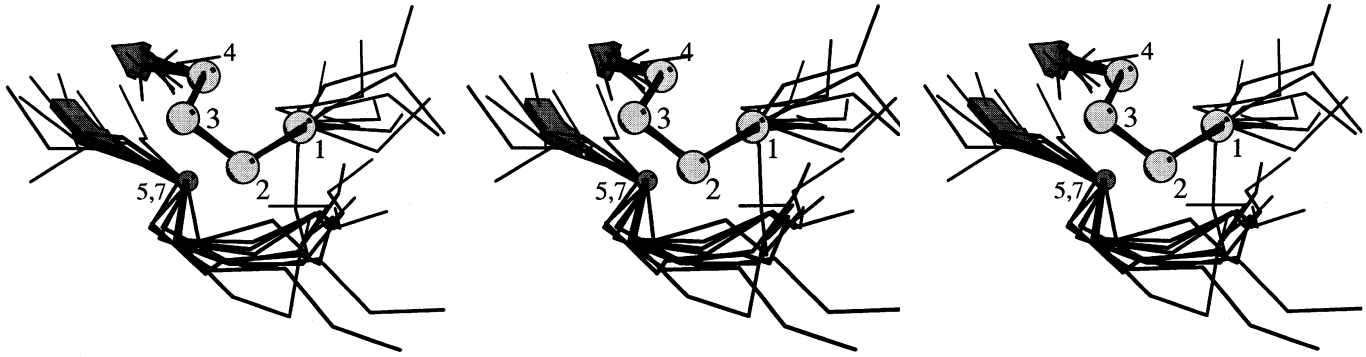

Fig. 7. Stereoscopic view of the local structures with NNT code 571234: 61857234: 0: 18723564: 0. The amino acid residues are shown in Table III (h). Ten structures are superposed. See also the caption of Figure 3.

β-bulge, which is an irregular region in a β-sheet lacking the normal pattern of hydrogen bonding. The type of the β-bulge shown here is called antiparallel (A) and G1 (G) according to Chan *et al.* (1993) (the type was defined by the program PROMOTIF [Hutchinson and Thornton (1996) in this study]). The three key residues, called X, 1 and 2, are $v_5 = v_7$, $v_3 = v_{15}$ and v_4, respectively. The fact that residue 1, i.e. $v_3 = v_{15}$, is Gly is a characteristic point for the β-bulge of type AG. It is also remarkable that the amino acid type of v_4 is Asp for seven of ten, and the remaining are Glu and Gln.

Ca^{2+} binding sites of the EF hand, i.e., loop structures connecting the E and F α-helices in Ca^{2+} binding proteins (Kretsinger, 1980) are found;

NNT code = 7182346: 0: 6178234: 7128354: 1823467,
$\{v_1–v_8\}$ = {i, i+3, i+4, i+5,, i+6, i–3, i+2}

(the data is not shown here). $N_{NNT} = 6$. The residues at vertices $v_{17} = v_{20}$, $v_8 = v_{14}$, v_3 and v_{19}, usually called X, Y, Z and –Y, respectively, provide oxygen atoms to chelate the Ca^{2+} ion. It is known that the Gly residue at v_4 is highly conserved in different Ca^{2+} binding proteins. The Asp and Asn residues at v_1, v_3 and $v_8 = v_{14}$, and the hydrophobic residues at $v_{13} = v_{16}$ are also well conserved.

Figure 8 and Table III (i) are an example related to structures including disulfide bridges. The fold of 2sn3 and 1pnh is classified as knottins (disulfide-bound fold containing β-hairpin with two adjacent disulfides), and that of 2tgi as cystine-knot cytokines (three disulfide links arranged in a knot-like topology) according to SCOP. 1ixa and 1hcc are also disulfide-rich small proteins. However, 1avdA and 1ppn are classified as rather different fold types than the others, although they also have disulfide bonds.

In 2sn3, 1ixa, 1pnh and 2tgi the Cys residues at v_5 and v_{17} form a disulfide bond, while the residues at these two site are not Cys in the other proteins. There are varieties of the formation of the disulfide bonds of the Cys residues at v_2. In the disulfide bonds 16–41 of 2sn3, 78–15 of 2tgi and 47–5 of 1hcc the counterpart Cys residues do not belong to this local structure of v_1 to v_{20}, while those in the disulfide bonds 62–51 of 1ixa and 83–4 of 1avdA are included in this local structure. There are other disulfide bonds: 8–26 of 1pnh and 44–109 of 2tgi. The local structure of 1ppn has no Cys residues, although this protein has three disulfide bonds in the other region. The fitting of the structures in Figure 8 is not good when compared with other examples shown in this paper. Nonetheless, these structures seem to have common characteristic features.

Discussion

In this paper we have proposed a novel method to detect a motif of local structures in protein conformations using the Delaunay tessellation. There are motifs at various levels, such as amino acid sequences, secondary structures, supersecondary structures, domains and tertiary structures. The motifs detected in this paper are parts of secondary and supersecondary structures, and other kinds of structures of similar size not directly related to the secondary structures. In this sense we call them motifs of local structures.

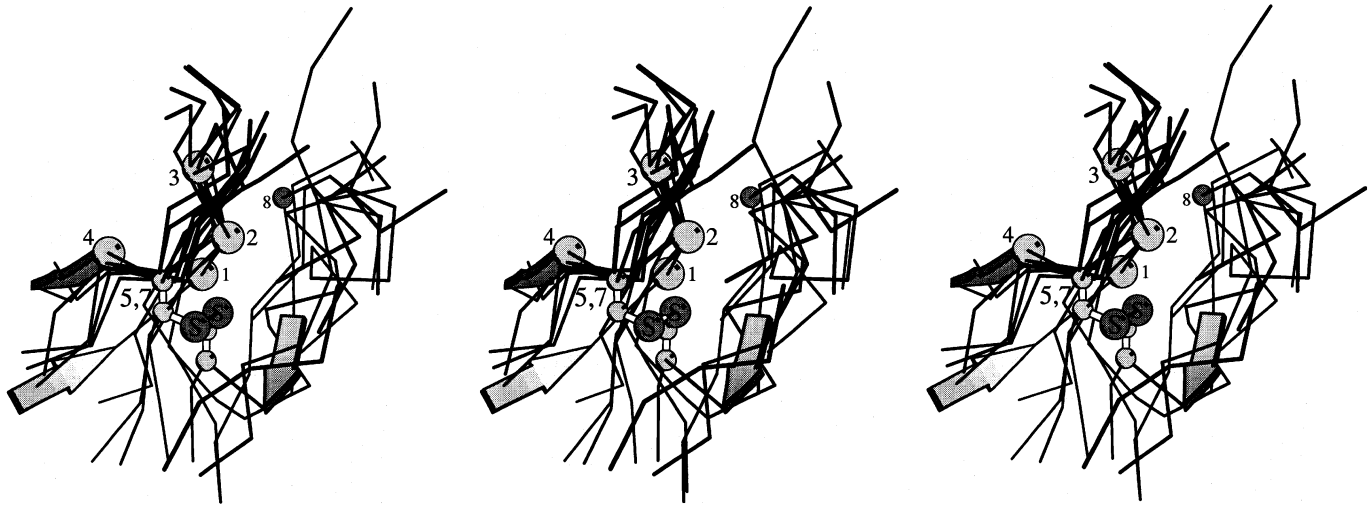

Fig. 8. Stereoscopic view of the local structures with NNT code 8123574: 67125834: 0: 86125734: 6182345. The amino acid residues are shown in Table III (i). Seven structures are superposed. See also the caption of Figure 3.

The Delaunay tessellation divides the protein interior into the Delaunay tetrahedra whose vertices are C_α atom positions uniquely. The edges of the tetrahedra can be regarded as an expression of the network of interactions between residues in 3D space. We intended to translate this 3D information about the network of the interactions into a one-dimensional sequence of digits, i.e., code, in this paper. We have achieved this by the unique numbering method of the vertices so that the code has information about the spatial arrangement of C_α atoms and the sequential connectivity of the chain.

The local structures with the same NNT code can be collected easily by computer. The structures with the same code resemble each other for most cases. Although the NNT code can include information about topographical structure of the residues, it cannot contain their 3D structure explicitly. Therefore, the fact that the local structures with the same code are similar to each other is not trivial. In fact, there are the cases where few exceptional structures are included even for the NNT codes. Consequently, superposition of the structures is necessary to assess the degree of resemblance. Nonetheless, the method of structural comparison based on topographical properties is important, if we consider the structural flexibility of protein molecules (Mizuguchi and Go, 1995).

The structures with the same code have a large variety in some cases (e.g., Figures 3, 5 and 8). Such a variety comes from the variety of relative spatial positions of the constituent segments and/or from the variety of structures of the segments in themselves. Nonetheless, the structures with the same code appear to have some common features. For example, favorable amino acid residues are limited at some particular vertices in Figure 3 and Table III (c) and (d), and the locations of disulfide bridges are common in Figure 8. β-α-β in Figure 5 and α-α on a β-strand in Figure 6 are typical supersecondary structures. The present method detects a motif of local structures from a topographical point of view rather than a superposition of the structures. This is the reason why such motifs can be detected by this method and one of the characteristic points of the method.

Although the final assessment of the degree of structural similarity should be carried out by their superpositions, a residue number pattern of the eight vertices v_1 to v_8 is useful for the first assessment. In most cases a given NNT code has only one residue number pattern, although there are some codes which correspond to more than one residue number patterns. A gap or insertion is usually responsible for this difference, but there are few cases that completely different structures corresponds to the same NNT code. Although there are some exceptions, the NNT code has enough ability to distinguish the local structures in general.

Since there are a huge number of NNT codes, it is impossible to see all local structures corresponding to every code on a graphic display at present. Currently we directed our attention to the most abundant codes and to the codes corresponding to the structures in which the amino acid types are likely to be limited at some particular vertices.

The most abundant codes are, as expected, related to α-helix and β-sheet. It should be emphasized, however, that the secondary structures are not taken into account explicitly in the Delaunay tessellation and the code assignment procedures. As far as the interactions between nearest neighbouring residues in space are taken into account, it is natural that these structures are identified as a motif. However, the motifs defined in this study differ from other motifs proposed up to now, even if our discussion is confined to the secondary structures. First, the local structures corresponding to the NNT codes are neither one whole α-helix nor one whole β-strand, but parts of the α-helix and the β-sheet. In addition some of the residues (i.e., structures they form) surrounding these parts are taken into account in the code. Consequently there are many codes related to the α-helix and the β-sheet reflecting the differences of structures located near the relevant secondary structures. More detailed classification of the α-helix and β-sheet based on such differences may be possible so that we can understand interactions between residues to stabilize these structures.

The local structures not related directly to the secondary structures are also detected as shown in the Miscellaneous motifs subsection of Results. As emphasized above, these structures can be detected, because we did not put any presupposition about secondary structures into the method. Again we want to emphasize that various kinds of motifs, such as those shown in Results, are detected in the same procedures, i.e., the Delaunay tessellation and code assignment to the Delaunay tetrahedron.

Representation of the local structure by the code is convenient in finding similar local structures by computer. At first the

Delaunay tessellation is applied to all proteins in the database, and then codes are assigned to the Delaunay tetrahedra obtained. The whole tetrahedra obtained for all the proteins in the database can be sorted easily with respect to the codes assigned, because the code is expressed as a number. As a consequence of the sorting, the tetrahedra, i.e., local structures, with the same code are collected automatically. We can use this set of sorted local structures as a new secondary database of protein structures.

This database can be used as follows. Suppose that we have a protein we wish to study. At first we can apply the Delaunay tessellation to this protein, and assign a code to every Delaunay tetrahedron obtained. Then, for each tetrahedron, i.e., for each local structure of this protein, we can easily derive the local structures with the same code from the database. Information about similar structures found in different proteins is helpful to understand and design them. However, it may be more helpful if we can know the different codes corresponding to similar local structures. In other words, clustering of the codes may be useful in such a database.

One of the difficulties of expressing local structures with a code is the problem of visualizing the local structures from that code. At present we have no idea about more comprehensive expression of the network of edges of the Delaunay tetrahedra. We may get more familiar with the codes in the future as the analyses of the codes progress.

There are huge number of codes. The analyses of local structures for whole codes have not been accomplished yet. Such analyses are now in progress.

References

Alexandrov,N.N. and Go,N. (1994) *Protein Sci.*, **3**, 866–875.
Alexandrov,N.N., Takahashi,K. and Go,N. (1992) *J. Mol. Biol.*, **225**, 5–9.
Aurora,R., Srinivasan,R. and Rose,G.D. (1994) *Science*, **264**, 1126–1130.
Barber,C.B., Dobkin,D.P. and Huhdanpaa,H.T. (1995) *ACM: Trans on Mathematical Software*, **22**, 469–483.
Chan,A.W.E., Hutchinson,E.G., Harris,D. and Thornton,J.M. (1993) *Protein Sci.*, **2**, 1574–1590.
Chelvanayagam,G., Roy,G. and Argos,P. (1994) *Protein Engng*, **7**, 173–184.
Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.
Flores,T.P., Orengo,C.A., Moss,D.S. and Thornton,J.M. (1993) *Protein Sci.*, **2**, 1811–1826.
Go,M. (1985) *Adv. Biophys*, **19**, 91–131.
Harper,E.T. and Rose,G.D. (1993) *Biochemistry*, **32**, 7605–7609.
Holm,L. and Sander,C. (1993) *J. Mol. Biol.*, **233**, 123–138.
Holm,L. and Sander,C. (1994) *Nucleic Acids Res.*, **22**, 3600–3609.
Holm,L. and Sander,C. (1996) *Methods Enzymol.*, **266**, 653–662.
Hutchinson,E.G. and Thornton,J.M. (1996) *Protein Sci.*, **5**, 212–220.
Johnson,M., Sutcliffe,M. and Blundell,T.L. (1990a) *J. Mol. Evol.*, **30**, 43–59.
Johnson,M., Sali,A. and Blundell,T.L. (1990b) *Methods Ezymol.*, **183**, 670–690.
Kline,T.P., Brown,K., Brown,S.C., Jeffs,P.W., Kopple,K.D. and Mueller,L. (1990) *Biochemistry*, **29**, 7805–7813.
Kobayashi,N., Yamato,T. and Go,N. (1997) *Proteins Struct. Funct. Genet.*, **28**, 109–116.
Kraulis,P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.
Kretsinger,R.H. (1980) *Crit. Rev. Biochem.*, **8**, 119–174.
Mizuguchi,K. and Go,N. (1995) *Protein Engng*, **8**, 353–362.
Munson,P.J. and Singh,R.K. (1997) *Protein Sci.*, **6**, 1467–1481.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,T. (1995) *J. Mol. Biol.*, **247**, 536–540.
Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.
Orengo,C.A., Mitchie,A., Jones,S., Jones,D., Swindells,M. and Thornton,J. (1996) *PDB Quarterly Newsletter*, **78**, 8–9.
Orengo,C.A. and Taylor,W.R. (1996) *Methods Enzymol.*, **266**, 617–635.
Richardson,J.S. (1981) *Adv. Protein Chem.*, **34**, 167–339.
Russell,R. and Barton,G. (1994) *J. Mol. Biol.*, **244**, 332–350.
Sali,A. and Blundell,T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.
Sander,C. and Schneider,R. (1991) *Proteins Struct. Funct. Genet.*, **9**, 56–68.
Schulz,G.E. and Schirmer,R.H. (1979) *Principles of Protein Structure*. Springer-Verlag, Berlin.
Singh,R.K., Tropsha,A. and Vaisman,I. (1996) *J. Comp. Biol.*, **3**, 213–221.
Taylor,W.R. and Orengo,C.A. (1989) *J. Mol. Biol.*, **208**, 1–22.
Vaisman,I.I., Brown,F.K. and Tropsha,A. (1994) *J. Phys. Chem.*, **89**, 5559–5564.
Voloshin,V.P., Naberukhin,Y.I. and Medvedev,N.N. (1989) *Mol. Simulation*, **4**, 209–227.
Yamato,T. (1996) *J. Mol. Graph.*, **14**, 105–107.
Yamato,T., Saito,M. and Higo,J. (1994) *Chem. Phys. Lett.*, **219**, 155–159.

Received April 3, 1998; revised May 26, 1998; accepted June 15, 1998

Appendix 1.

The following 293 proteins are used for the analysis in this paper. They have less than 25% sequence homology with each other:

129l, 1aaf, 1aaj, 1aapA, 1aba, 1abk, 1abmA, 1add, 1ads, 1aep, 1alkA, 1aozA, 1apa, 1apmE, 1aps, 1arb, 1atnA, 1atx, 1avdA, 1avhA, 1baa, 1babB, 1bbl, 1bbpA, 1bbt1, 1bbt2, 1bgeB, 1bgh, 1bllE, 1bovA, 1brd, 1bsaA, 1btc, 1bw3, 1c5a, 1caj, 1cauB, 1ccr, 1cdb, 1cde, 1cdtA, 1chrA, 1cid, 1cmbA, 1cobA, 1colA, 1cpcA, 1cpcL, 1cpt, 1crl, 1cseI, 1ctaA, 1d66A, 1dfnA, 1dhr, 1dog, 1dsbA, 1eaf, 1eco, 1ede, 1erp, 1ezm, 1faiL, 1fas, 1fbaA, 1fc1A, 1fc2C, 1fdd, 1fha, 1fiaB, 1fod4, 1fxiA, 1gal, 1gatA, 1gdhA, 1gky, 1glaF, 1glaG, 1glt, 1gmfA, 1gox, 1gps, 1gsrA, 1hbq, 1hcc, 1hddC, 1hdxA, 1hgeB, 1hivA, 1hleA, 1hleB, 1hmy, 1hra, 1hsbA, 1huw, 1ipd, 1isuA, 1ixa, 1lcdA, 1le4, 1lenA, 1lgaA, 1lis, 1ltsA, 1ltsC, 1ltsD, 1mdaA, 1mdc, 1mioC, 1mrt, 1mup, 1mypC, 1nar, 1ndk, 1nipB, 1nrcA, 1nxb, 1ofv, 1omb, 1omf, 1omp, 1onc, 1osa, 1pda, 1pdc, 1pdgB, 1pfkA, 1phh, 1plc, 1pnh, 1poa, 1poc, 1poxA, 1ppfE, 1ppn, 1ppt, 1prcC, 1prcM, 1pyp, 1r094, 1r1a2, 1rcb, 1rec, 1rhd, 1ribA, 1rinB, 1rnd, 1rprA, 1rveA, 1s01, 1sbp, 1sgt, 1shaA, 1shfA, 1sltA, 1smrA, 1snc, 1spa, 1sryA, 1tabI, 1tbpA, 1ten, 1tfd, 1tfi, 1tgsI, 1tie, 1tlk, 1tml, 1tnfA, 1tplA, 1tpm, 1trb, 1troA, 1ttbA, 1ula, 1utg, 1vaaB, 1vil, 1vsgA, 1wsyA, 1wsyB, 1zaaC, 2aaiB, 2achB, 2atcB, 2ayh, 2azaA, 2bbkH, 2bds, 2bopA, 2bpa1, 2bpa2, 2bpa3, 2cas, 2cbh, 2ccyA, 2cdv, 2cmd, 2cp4, 2cpl, 2crd, 2cro, 2ctc, 2cts, 2cyp, 2dnjA, 2dri, 2ech, 2end, 2er7E, 2gb1, 2hipA, 2hpdA, 2ihl, 2lh2, 2liv, 2mev1, 2mhr, 2mnr, 2mrb, 2ms2A, 2msbA, 2mtaC, 2pcdA, 2pf1, 2pgd, 2pia, 2plv1, 2pmgA, 2por, 2reb, 2rn2, 2sas, 2scpA, 2sga, 2sim, 2sn3, 2snv, 2spo, 2stv, 2tbvA, 2tgi, 2tmvP, 2tscA, 2ztaA, 3adk, 3b5c, 3chy, 3cla, 3cox, 3dfr, 3egf, 3gapA, 3gbp, 3grs, 3inkC, 3monA, 3pgk, 3rubS, 3sdhA, 3sgbI, 3tgl, 4blmA, 4cpaI, 4enl, 4fgf, 4fxn, 4gcr, 4htcI, 4insB, 4rcrH, 4sbvA, 4sgbI, 4tgf, 4ts1A, 4xis, 4znf, 5fbpA, 5nn9, 5p21, 5timA, 5znf, 6taa, 7apiB, 8abp, 8atcA, 8catA, 8i1b, 8rxnA, 9ldtA, 9rnt, 9rubB, 9wgaA.