Local feature detection in protein and RNA aptamer structures based on differential geometry of tetrahedron sequences

Naoto Morikawa

August 30, 2007

Abstract

Local protein structure analysis is informative to protein structure analysis and has been used successfully in protein structure prediction and others. Proteins have recurring structural features, such as helix caps and beta turns, which often have strong amino acid sequence preferences. And challenges for local structure analysis have been identification and assignment of such common short structural motifs.

In this paper, we introduce a new differential geometrical approach for local structure analysis, in which local conformations of protein backbones are described using variation of gradient along folded tetrahedron sequences. Using the approach, we could capture recurring local structural features without any structural motifs, which makes local structure analysis more accurate and objective. At the same time, it makes local structure analysis simpler since one does not need to identify and assign structural motifs at all.

The programs and examples are available from `http://www.genocript.com`.

Keywords: *local protein structure; one-dimensional representation; structural motif; tetrahedron sequence; discrete differential geometry*

Email: nmorika@genocript.com

1 Introduction

Local protein structure analysis is informative to protein structure analysis (Unger & Sussman, 1993; Rooman *et al.*, 1990; Kolodny *et al.*, 2002) and has been used successfully in protein structure prediction (Baystroff *et al.*, 2000; Hunter & Subramaniam, 2003; Sander *et al.*, 2006; Benros *et al.*, 2006), fold recognition, remote homolog detection, and others.

Proteins have recurring structural features, such as helix caps and beta turns, which often have strong amino acid sequence preferences. Challenges for local structure analysis have been identification and assignment of such common short structural motifs. Identification involves description of protein backbone conformation and definitions of the structural motifs. And assignment is not a trivial task, due to the variations observed in nature when compared to ideal ones.

For example, Unger & Sussman (1993) have identified about 100 structural motifs by clustering a set of amino acid fragments of length six based on structural similarity. As a measure of similarity, they used the root mean square deviation (RMSD) computed on the six α-carbons for each pair of fragments. Rooman *et al.* (1990) obtained 16 structural motifs of different length by clustering a set of fragments of length seven into four groups (α-helix, β-strand, turn, and coil) using also the RMSD on α-carbons. And de Brevern *et al.* (2000) proposed 16 structural motifs of length five, called Protein Blocks, which were obtained by a cluster analyzer based on the self-organized maps. They used the RMSD of backbone dihedral angles (ϕ, ψ) as a similarity measure.

On the other hand, Bystroff & Baker (1998) considered clustering of amino acid fragments based on sequence similarity to identify 82 clusters of different length, where each cluster represents a sequence neighborhood that adopt only one or a few local structures. They used a combination of α-carbon distance and backbone dihedral angles to characterize the structural motifs. And Sander *et al.* (2006) combined both of structure-based and sequence-based clustering to identify 27 structural motifs of length seven.

In this paper, we introduce a new differential geometrical approach for local structure analysis. As for differential geometrical description of protein structures, the early work of Rackovsky & Scheraga (1978) described protein backbones as broken lines, where each line corresponds to the virtual-bond between consecutive α-carbons. In contrast, we describe local conformation of protein backbones using variation of gradient, i.e., the second derivative, along folded tetrahedron sequences. Then, we could capture recurring local structural features without any structural motifs, which makes local structure analysis more accurate and objective. At the same time, it makes local structure analysis simpler since one does not need to identify and assign structural motifs at all.

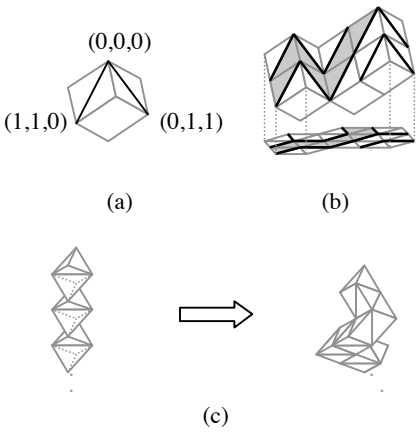

Figure 1: Basic ideas. (a): Division of facets of unit cube $[0,1]^3$. (b): Division of the surface of "peaks and valleys" of cubes and projection of the surface on a hyperplane. (c): Folding of a tetrahedron sequence.

2 Differential geometry of tetrahedron sequences

2.1 Basic ideas

Let's consider unit cube $[0,1]^3$ in the three-dimensional Euclidean space $\mathbf{R}^3$ and divide each of the three facets which contain $(0,0,0)$ into two triangles along diagonal, as shown in Figure 1 (a). Then, if we pile the cubes up in the direction of $(-1,-1,-1)$, we would obtain "peaks and valleys" of the cubes, where the facet division of each cube makes up a surface division of the "peaks and valleys" (Figure 1 (b) above). And a "flow" of triangles in $\mathbf{R}^2$ is obtained by projecting the surface onto a hyperplane, as shown in Figure 1 (b) below. For example, grey "slant" triangles on the surface (above) specify the trajectory of grey "flat" triangles on the hyperplane (below).

Similarly we obtain a "flow" of tetrahedrons in $\mathbf{R}^3$ by considering "peaks and valleys" of 4-cubes in $\mathbf{R}^4$. In this case, each trajectory of tetrahedrons could be obtained by folding a tetrahedron sequence which satisfies the following conditions (Figure 1 (c)) : (i) Each tetrahedron consists of four short edges and two long edges, where the ratio of the length is $\sqrt{3}/2$ and (ii) Successive tetrahedrons are connected via a long edge and have a rotational freedom around the edge.

In particular, we could compute the differential structure on a trajectory without considering 4-cubes. In the following, we describe the differential geometry of tetrahedron sequences using the tetrahedron sequence specified above. For mathematical foundation, see Morikawa (2006).

2.2 Gradient of tetrahedrons

The *gradient* of a tetrahedron on a trajectory is the direction of the short edge which is not shared with the adjacent tetrahedrons on the trajectory.

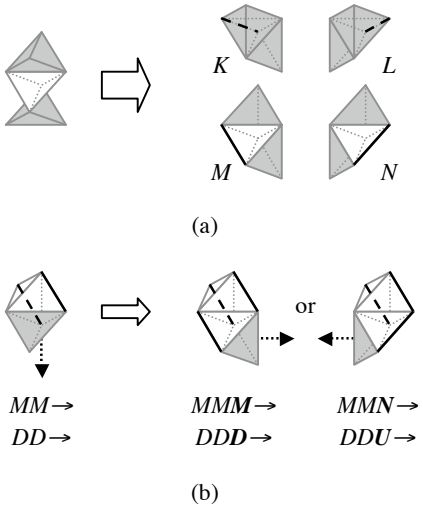

Figure 2: Differential structure of a tetrahedron sequence. (a): Gradient of a tetrahedron. The gradient of a tetrahedron (white) is determined by the spatial arrangement of adjacent tetrahedrons on the trajectory. And each tetrahedron assumes one the four values, K, L, M, and N as its gradient. (b): Variation of gradient along a trajectory. The next tetrahedron (grey in the left figure) could assume one of the two values shown on the right as its gradient. The character strings show the corresponding sequences of gradients (upper) and second derivatives (lower). Note that the position of the new next tetrahedron (grey in the right figure) is also determined when gradient is assigned.

As an example, let's consider a tetrahedron sequence of length three (Figure 2 (a) left). Then, we obtain four types of trajectories shown on the right by folding the sequence. As you see, each trajectory specifies a short edge (bold line) of the white tetrahedron which is not contained in the adjacent grey tetrahedrons. In the following, we use four alphabets K, L, M, N to denote the direction of the short edges. In particular, the gradient of a tetrahedron is specified using the four letters.

For example, let's consider the tetrahedron sequence shown in Figure 2 (b) left. The gradient of the current and previous tetrahedrons (white) are M and the gradient of the next tetrahedron (grey) is not yet determined. Note that the gradient of the next tetrahedron (grey) is either M or N because of the restriction of the sequence structure, as shown in Figure 2 (b) right.

2.3 Variation of gradient along trajectory

Now let's consider variation of gradient along a trajectory, i.e., the second derivative of tetrahedrons. As we see above, because of the restriction of the sequence structure, each tetrahedron could assume only one of two gradient values which are determined by the gradient of the preceding tetrahedron. Thus, we could describe variation of gradient using a binary sequence of, say, U and D. The computing rule is simple;

change value if gradient changes.

As an example, let's consider the tetrahedron sequence shown in Figure 2 (b) left again. Suppose that the second derivative of the current tetrahedron (white) is D. Then, the second derivative of the previous tetrahedron (white) is also D since they have the same gradient. As for the next tetrahedron (grey), there are two possibilities (Figure 2 (b) right). In the left case, the second derivative of the next tetrahedron is also D since it has the same gradient as that of the current one. In the right case, the second derivative of the next tetrahedron is U since gradient changes from M to N. In either case, we have obtained a binary sequence of length three, DDD or DDU, which describes the shape of the corresponding trajectory.

In the next section, we will use variation of gradient to encode the local structure of proteins and RNA aptamers.

3 Methods

3.1 Broken line approximation of protein and RNA structures

Upon approximation of protein structures, positions of α-carbon atoms only are considered. That is, we identify amino acids with their α-carbon atoms and study the structure of the broken line obtained by connecting the α-carbon atoms of adjacent amino acids.

As for RNA molecules, positions of C1'-carbon atoms only are considered. And we study the structure of the broken line obtained by connecting the adjacent C1'-carbon atoms of adjacent nucleic acids.

In the following, broken lines are approximated by folded tetrahedron sequences to detect their local features, where we permit translation and rotation during the folding process since we would often obtain a more complicated trajectory than the original broken line by simple folding.

3.2 5-tile coding of local structures

To study local structure of a protein, we consider all the amino acid fragments of length *five* occurred in the protein. (It will turn out that length five is enough to detect local features.)

Each fragment is approximated by a folded tetrahedron sequence of length five, starting from the middle point amino acid, say A. And variation of gradient along the sequence is computed to encode its structural features. We call the resulting $\{D, U\}$-valued sequence of length five the *5-tile code* of A.

In the case of RNAs, we consider all the nucleic acid fragments of length five.

3.3 Encoding algorithm

Now we will explain the algorithm of tetrahedron folding with translation and rotation. As an example, let's consider the amino acid fragment $AA[-2]$-$AA[-1]$-$AA[0]$-

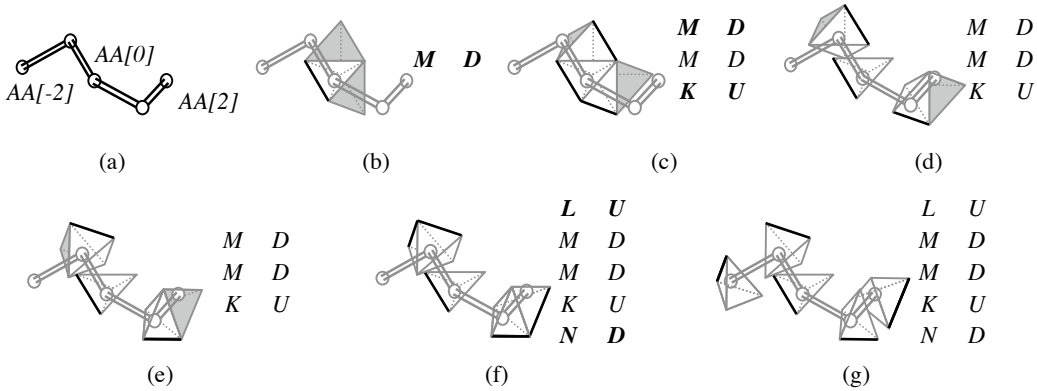

Figure 3: Algorithm of the 5-tile coding. (a): Broken line which represents the structure of amino acid fragment $AA[-2]$-$AA[-1]$-$\cdots$-$AA[2]$ to be encoded. (b): Step 1. (c): Step 2. (d): Step 3. (e): Step 4. (f): Step 5. (g) Step 6. The character strings show the corresponding sequences of gradients (left) and second derivatives (right), where top are the those of $T[-2]$ and bottom are those of $T[2]$.

$AA[1]$-$AA[2]$ of figure 3 (a) and compute the 5-tile code of $AA[0]$ using a sequence of five tetrahedrons $T[-2]$-$T[-1]$-$T[0]$-$T[1]$-$T[2]$.

3.3.1 Step 1

Align tetrahedron $T[0]$ (white) with amino acid $AA[0]$ and set initial values (Figure 3 (b)). In this example, the gradient and second derivative of $T[0]$ is M and D respectively.

Then, the initial positions of adjacent tetrahedrons $T[\pm 1]$ (grey) are also determined, which are moved to the positions of $AA[\pm 1]$ respectively later.

3.3.2 Step 2

Assign gradient to adjacent tetrahedrons $T[\pm 1]$ considering the direction of $AA[\pm 2]$ respectively (Figure 3 (c)). For example, tetrahedron $T[1]$ could assume M or K as its gradient. And the next tetrahedron (grey) becomes closer to $AA[2]$ if K is assumed. Thus, the gradient of $T[1]$ is K and its second derivative is U since the gradients of $T[0]$ and $T[1]$ are different. In the same way, $T[-1]$ is assigned M and D as its gradient and second derivative respectively.

Note that the initial positions of adjacent tetrahedrons $T[\pm 2]$ (grey) are also determined, which are moved to the positions of $AA[\pm 2]$ respectively later.

3.3.3 Step3

Translate tetrahedrons $T[\pm 1]$ to the positions of $AA[\pm 1]$ respectively (Figure 3 (d)). Adjacent tetrahedrons $T[\pm 2]$ (grey) are also moved with $T[\pm 1]$ respectively.

3.3.4 Step4

Rotate tetrahedrons $T[\pm 1]$ at the positions of $AA[\pm 1]$ so that the bold edges become parallel to the direction from $AA[0]$ to $AA[\pm 2]$ respectively (Figure 3 (e)). Adjacent tetrahedrons $T[\pm 2]$ (grey) are also moved with $T[\pm 1]$ respectively.

3.3.5 Step5

Assign gradient to adjacent tetrahedrons $T[\pm 2]$ considering the direction of $AA[\pm 2]$ respectively (Figure 3 (f)). For example, tetrahedron $T[2]$ could assume N or K as its gradient. And the next tetrahedron (not shown) becomes closer to $AA[2]$ if N is assumed. Thus, the gradient of $T[2]$ is N and its second derivative is D since the gradients of $T[1]$ and $T[2]$ are different. In the same way, $T[-2]$ is assigned L and U as its gradient and second derivative respectively.

3.3.6 Step6

Translate tetrahedrons $T[\pm 2]$ to the positions of $AA[\pm 2]$ respectively (Figure 3 (g)). And we have obtained binary sequence $UDDUD$, the 5-tile code of $A[0]$, which describes the shape of the amino acid fragment shown in Figure 3 (a).

3.4 One-letter representation of 5-tile codes

To save space, we use numerals and alphabets to denote 5-tile code $C_1C_2C_3C_4C_5$. First, compute the value Y of the code which is defined as follows: $Y = 2^4 * C'_1 + 2^3 * C'_2 + 2^2 * C'_3 + 2 * C'_4 + C'_5$, where $C'_i = 1$ if C_i is equal to U and $C'_i = 0$ if not. Then, assign the number to the code if the value Y is less than 10. Otherwise, assign the $(Y - 9)$-th alphabet to the code.

For example, $DDDUU$ corresponds to binary number 00011 and $Y = 3$. Thus, 3 is assigned to the code. On the other hand, $DUDUD$ corresponds to binary number 01010 and $Y = 10$. Thus, the first alphabet A is assigned to the code.

4 Results

In the following, structures of proteins and RNAs in the PDB database are encoded into 5-tile code sequences. For comparison, assignment of DSSP state (Kabsch & Sander, 1983) and the Protein Blocks (de Brevern *et al.*, 2000) are also considered.

The DSSP program computes secondary structure, hydrogen bonding, and solvent exposure of a protein, to assign one of eight states to each residue, where E corresponds to strand, H, G, I to helix, T to turn, S to bend, B to bridge, and $-$ to no assigned structure.

On the other hand, the Protein Blocks (PBs) are a set of short structural motifs composed of sixteen folding patterns of five consecutive α-carbons. Protein backbone structures are approximated by the blocks based on the backbone diehedral angles, where D corresponds to strand, M to helix, K, L to N-cap of helix, N, O to C-cap of helix, and K, L to N-cap alpha, G, H, I, J to mainly coil, and so on.

```
MISDEQLNSLAITFGIVM***YHAVDSTMSPKN
--000RQAAAAAAAHAAA***AAB0R00000--
--SS-GGGHHHHHHHHHH***HHHHTS------
--JACKLMMMMMMMMMM***MMMMMMOCDF--
```

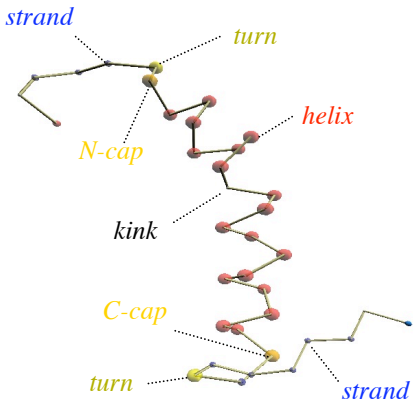

Figure 4: Structural features of 1RKL and spatial arrangement of its 5-tile codes. The character strings show the amino acid sequence, the 5-tile code sequence, DSSP assignment, and PB sequence of 1RKL (from top to bottom).

Programs and examples of the 5-tile coding are available from http://www.genocript.com.

4.1 Structural features of 1RKL (protein)

To examine the capability of 5-tile codes to detect local structural features, let's consider transferase 1RKL as an example. Figure 4 shows the structural features of 1RKL and spatial arrangement of its 5-tile codes, where blue small balls stand for 0, yellow balls stand for R, orange balls for B or Q, and red balls for A. As you see, there exist clear correspondences between local features and 5-tile codes. That is, correspondences between strand and 0, turn and R, N-cap and Q, helix and A, and C-cap and B. Moreover, one could identify a kink in the helix as a break of consecutive As (H in this example).

On the other hand, DSSP considers N-terminal of the helix to be 3-helix G and assigns no structure to the N-terminal (amino acid E) of the 3-helix. Furthermore, it considers that the helical structure H continues until the turn T on amino acid D, omitting the strand between amino acid A and D.

As for the PB sequence, it dose not capture the delicate structure around the C-terminal of the helix. And it is difficult to interpret its description of the strands at both ends, i.e., JAC and CDF.

Finally, both the DSSP state and PB sequences fail to capture the kink in the helix.

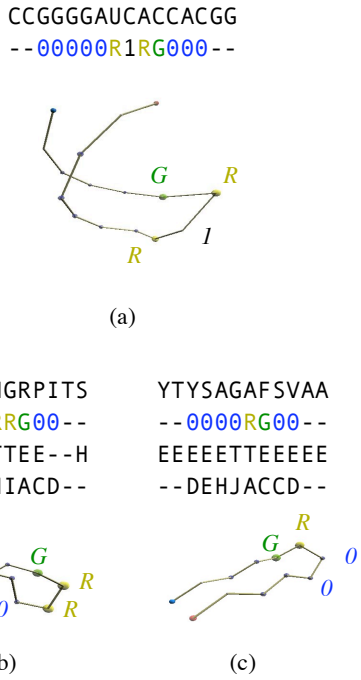

Figure 5: Structural motifs. (a): An RNA aptamer (1U1Y). The character strings in (a) show the nucleic acid sequence (upper) and 5-tile code sequence of 1U1Y. (b): Type-I' beta turn (2CPP residue 224-235). (c): Type-II' beta turn (2POR residue 121-132). The character strings in (b) and (c) show the amino acid sequence, the 5-tile code sequence, DSSP assignment, and PB sequence of the turns (from top to bottom).

4.2 Structural motifs: beta turns and 1U1Y (RNA aptamer)

As for structural motifs, we could also detect them using 5-tile code sequences. Figure 5 (b) and (c) show type-I' and type-II' beta turns which correspond to 5-tile sequences, $0RRG$ and $00RG$ respectively. As you see, the difference of the two turns is captured as the difference of the structure around the second amino acid of the turn, i.e., R and 0.

On the other hand, DSSP couldn't distinguish the turns from one another since both turns have the same DSSP state sequence, $ETTE$. And the PB sequences of the turns, $EHIA$ and $HJAC$, shows no sign of structural similarity between them.

Finally, note that one could also compare local structure of proteins with that of RNA aptamers using the 5-tile encoding. RNA aptamers are nucleic acid sequences which can recognize target molecules with high affinity and specificity. And they can be used as macromolecular drugs, in substitution for antibodies.

For example, figure 5 (a) shows RNA aptamer 1U1Y, which folds into a beta turn-like structure whose 5-tile sequence is $R1RG$. As you see, the first half structure of the turn is different from that of type-I' and type-II' beta turns and its 5-tile code sequence,

Table 1: Frequency distribution of the occurrence of the 5-tile codes, DSSP states, and PBs in nine superfolds (1THB chain A and B, 256B chain A, 1APS, 1UBQ, 2FOX, 7TIM chain A, 1I1B, 2BUK, and 2RHE).

	5-tile code		PB		DSSP	
	code	*(%)*	*code*	*(%)*	*code*	*(%)*
1	0	(40.1)	M	(35.1)	H	(34.5)
2	A	(32.6)	D	(16.5)	E	(24.0)
3	R	(7.7)	C	(8.2)	-	(15.4)
4	Q	(6.8)	F	(5.7)	T	(11.0)
5	B	(6.0)	L	(5.4)	S	(9.0)
6	G	(5.0)	K	(5.2)	G	(4.9)
7	O	(0.4)	A	(4.3)	B	(1.2)
8	H	(0.4)	P	(3.6)		
9	3	(0.3)	H	(3.0)		
10	1	(0.3)	E	(2.6)		
11	9	(0.1)	B	(2.6)		
12	8	(0.1)	I	(2.3)		
13	I	(0.1)	O	(2.0)		
14			N	(1.6)		
15			G	(1.2)		
16			J	(0.7)		

$R1$, is different from that of the others accordingly. We will show below that the structural feature around nucleic acid C, i.e., 5-tile code *1*, is one of the characteristics of RNA aptamer structures.

4.3 Occurrence of 5-tile codes in nine superfolds

Superfolds are protein folds known to recur in proteins, having neither sequence nor functional similarity (Orengo *et al.*, 1994). And table 1 shows the frequency distribution of the occurrence of the 5-tile codes, DSSP states, and PBs in nine superfolds: 1THB chain A and B, 256B chain A, 1APS, 1UBQ, 2FOX, 7TIM chain A, 1I1B, 2BUK, and 2RHE (identical chains are discarded).

In the case of 5-tile codes, about 40% are *0* and the top six 5-tile codes cover 98.2% of all the amino acids. The rest of the code often indicate structural distortion such as kinks in helices, as shown in figure 4. Another characteristic is the imbalance between the frequencies of occurrence of G ($= UDDDD$) and *1* ($= DDDDU$).

In the case of PBs, D occupies 16.5% and the top six PBs cover only 76.1%. That is, PBs are more evenly distributed over amino acids than 5-tile codes.

As for DSSP states, E covers 24.0%, more than the frequency of PB block D and less than that of 5-tile code *0*. And 15.4% have not assigned any structure.

Table 2: DSSP assignment over the amino acids of a 5-tile code in nine superfolds (1THB, 256B, 1APS, 1UBQ, 2FOX, 7TIM, 1I1B, 2BUK, and 2RHE), where helix corresponds to DSSP states H, G, I, strand to E, turn to T, bend to S, bridge to B, and else to $-$. For example, 2% of the amino acids of 5-tile code *0* are helix, 56% are sheet, and so on (top). And 35% of all the amino acids are helix, 27% are sheet, and so on (bottom).

			Frequency distributions of DSSP assignment (%)					
Code	*Occur.*	*(%)*	*Helix*	*Strand*	*Turn*	*Bend*	*Bridge*	*Else*
0 (=DDDDD)	534	(44.0)	2	56	4	8	2	27
A (=DUDUD)	348	(28.6)	97	0	3	0	0	0
R (=UUDUU)	100	(8.2)	0	8	34	30	1	27
Q (=UUDUD)	81	(6.7)	51	0	38	11	0	0
B (=DUDUU)	73	(6.0)	40	0	45	11	0	4
G (=UDDDD)	63	(5.1)	3	25	6	37	5	24
Else	16	(1.3)	25	13	25	25	0	13
All	1215	(100.0)	35	27	12	10	1	16

Table 2 shows the DSSP assignment over the amino acids of a 5-tile code in the nine superfolds (all chains are included). According to the table, 56% of the amino acids of 5-tile code *0* are assigned strand (E) and most of the strands are assigned 5-tile code *0* with some exceptions which are assigned G or R. Note that only 35% of all the amino acids are assigned strand but 44% are assigned 5-tile code *0*.

As for helices, 97% of 5-tile code A are assigned helix (H, G, or I). And 51% of Q and 40% of B are also assigned helix, which indicates DSSP makes helices longer than the 5-tile coding. Note that only 28.6% of all the amino acids are assigned 5-tile code A but 35% are assigned helix.

As for turns and bends, 64% of the amino acids of 5-tile code R are assigned turn (T) or bend (S). Inversely, the 5-tile codes of the turns are distributed over R, Q, or B evenly and that of bends are distributed over R, G, Q, or B. In particular, 5-tile codes describe more details about turn/bend structures than DSSP states.

4.4 Frequency distribution of 5-tile codes

ASTRAL SCOP 1.71 (95%) sequences are amino acid sequences with less than 95% identity, extracted from the SCOP database (Chandonia *et al.*, 2004). And table 3 shows the frequency of occurrence of the 5-tile codes in the ASTRAL SCOP sequences and 33 RNA aptamers registered in the PDB database (almost all the RNA aptamers in the PDB).

First of all, 41.2% of the amino acids in ASTRAL SCOP sequences are assigned

Table 3: Frequency of the occurrence of the 5-tile codes in ASTRAL SCOP 1.71 (95 %) sequences and 33 RNA aptamers registered in the PDB database.

	ASTRAL SCOP		RNA aptamer	
Code	*Occur.*	*(%)*	*Occur.*	*(%)*
0	898516	(41.2)	604	(71.0)
1	18364	(0.8)	23	(2.7)
2	970	(0.0)	5	(0.6)
3	6223	(0.3)	31	(3.6)
8	2033	(0.1)	1	(0.1)
9	1034	(0.1)	0	(0.0)
A	639440	(29.3)	0	(0.0)
B	125984	(5.8)	8	(0.9)
G	114560	(5.3)	32	(3.8)
H	6507	(0.3)	3	(0.4)
I	792	(0.0)	2	(0.2)
J	2369	(0.1)	14	(1.6)
O	9209	(0.4)	19	(2.2)
P	2027	(0.1)	4	(0.5)
Q	139088	(6.4)	9	(1.1)
R	212275	(9.7)	96	(11.3)

5-tile code *0* and 29.3% are assigned A. But the RNA aptamers have no nucleic acid of 5-tile code A and 71.0% of the nucleic acids are assigned 5-tile code *0*. In particular, the RNAs has no helix and the occurrence of B and Q are also low accordingly since they are usually correspond to caps of helices.

Instead of helix-related structures, 5-tile codes *1* ($= DDDDU$), *3* ($= DDDUU$), and O ($= UUDDD$) occur as well as R and G in the RNAs. In particular, *1* occurs as often as G and there even exists a short motif of 5-tile code sequence $1RG$, as shown in figure 5 (a).

As for proteins, they usually don't form neither slow curves of 5-tile code 3 or O, nor sharp turns of 5-tile code sequence $1RG$.

5 Discussion

Firstly, the 5-tile coding could detect both new local structures and structural distortions since 5-tile codes are computed directly from atomic coordinates. On the other hand, structural motifs are often identified by clustering a set of representative protein fragments, using unsupervised machine learning. Thus, they could not recognize new structures nor distortions.

Secondly, we could use 5-tile code sequences to compare protein structures. For

example, we have compared the structure of type-I' beta turn with that of typr-II' above, i.e., $0RRG$ and $00RG$. Note that, unlike the root mean square deviation (RMSD), we could identify the location where the two structures are different. For example, the type-I' and typr-II' turns of figure 5 are 75% identical and they differ on the second place. We have also detected the location of a kink in a helix successfully above (figure 4).

In particular, the 5-tile coding gives an example of quantitative similarity measures which is essential for a critical assessment of the quality of protein structure prediction. For example, one could locate the precise position of miss prediction by comparing the 5-tile code sequences of the predicted and the actual structures.

Moreover, comparison of 5-tile code sequences gives a fast and accurate search method to extract proteins of similar structure from a database since the computation is far more straightforward than that of the RMSD. See Koehl (2001) for more about similarity measures.

Thirdly, we could also describe structural similarity between different 5-tile codes objectively, using their U/D sequences. For example, 9 $(= DUDDU)$ is different from A $(= DUDUD)$ in the last two places. On the other hand, structural similarity between different structural motifs are rather ambiguous.

Finally, note that we could deal with structures of proteins and RNAs in the same framework. For example, we have compared the structure of an RNA aptamer with that of beta turns to identify their differences above (figure 5). And we have obtained new characterizations of protein structures by comparing the structures of proteins and RNAs, i.e., the "lack" of 5-tile codes 1, 3, and O.

References

[1] Benros, C., de Brevern, A. G., Etchebest, C. & Hazout, S. 2006 Assessing a novel approach for predicting local 3D protein structures from sequence, *Proteins* **62(4)**, 865-880.

[2] Bystroff, C. & Baker, D. 1998 Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* **281**, 565-77.

[3] Bystroff, C., Thorsson,V. & Baker, D. 2000 HMMSTR: A hidden Markov model for local sequence-structure motif. *J. Mol. Biol.* **281**, 565-577.

[4] de Brevern, A.G., Etchebest, C. & Hazout, S. 2000 Bayesian Probabilistic Approach for Predicting Backbone Structures in Terms of Protein Blocks. *Proteins* **41**, 271-287.

[5] Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. 2004 The ASTRAL compendium in 2004. *Nucleic Acids Research* **32**, D189-D192.

[6] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. & Sigrist, C. J. A. 2006 The PROSITE database. *Nucleic Acids Research* **34**, D227-D230.

[7] Hunter, C. & Subramaniam, S. 2003 Protein local structure prediction from sequence. *Protein* **50(4)**, 572-579.

[8] Kabsch, W. & Sander, C. 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.

[9] Koehl, P. 2001 Protein structure similarities. *Curr. Opin. Struct. Biol.* **11**, 348-353.

[10] Kolodny, R., Koehl, P., Guibas, L. & Levitt, M. 2002 Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.* **323**, 297-307.

[11] Morikawa, N. 2006 Discrete differential geometry of triangle tiles and algebra of closed trajectories. ArXiv: math.CO/0607051.

[12] Orengo, C.A., Jones, D.T. & Thornton, J.M. 1994 Protein superfamilies and domain superfolds. *Nature* **372**, 631-634.

[13] Petersen, T. N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G. P. & Lund, O. 2000 Prediction of protein secondary structure at 80% accuracy. *Proteins* **41**, 17-20.

[14] Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. 2002 Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47(2)**, 228-235.

[15] Porter, C. T., Bartlett, G. J. & Thornton, J. M. 2004 The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research* **32**, D129-D133.

[16] Rackovsky, S. & Scheraga, H.A. 1978 Differential Geometry and Polymer Conformation. 1. *Macromolecules* **11**, 1168-1174.

[17] Rooman, M., Rodriguez J. & Wodak, S. 1990 Automatic definition of recurrent local structure motifs in proteins. *J. Mol. Biol.* **213(2)**, 328-336.

[18] Sander, O., Sommer, I. & Lengauer, T. 2006 Local protein structure prediction using discriminative models, *BMC Bioinformatics* **7**, 14-26.

[19] Unger, R. & Sussman, J. L. 1993 The importance of short structural motifs in protein structure analysis. *J. Comput. Aided Mol. Des.* **7**, 457-472.