Representation of a 3D Protein Structure Using a Sequence of Line Segments
(Extended Abstract)

Tatsuya Akutsu and Hiroshi Tashimo

Department of Computer Science, Gunma University
1-5-1 Tenjin, Kiryu, Gunma 376 Japan
e-mail: {akutsu, tashimo}@keim.cs.gunma-u.ac.jp

1 Introduction

Comparing tertiary (or 3D) structures of proteins is very important in bio-informatics [2, 3, 4, 5]. In most of previous studies, a tertiary structure has been represented by a sequence of points or a list of types of fragments (for example, (α-helix, turn, α-helix, β-strand, $\cdots$)). However, either is not adequate in some cases since the former representation is too detailed while the latter representation is too rough. Thus, intermediate representation is sometimes required. In this extended abstract, we propose a new method for such representation, in which a 3D structure is represented by a sequence of line segments. Note that details are omitted here, and will be presented elsewhere.

2 The method

We assume that each 3D protein structure is stored as a sequence of points (i.e., a sequence of Cα atoms). Thus, we let $P = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_n)$ be a 3D structure, where each $\boldsymbol{p}_i$ denotes a point in the 3D Euclidean space. Then, we compute a sequence of line segments in the following way.

(i) Compute a sequence of lines approximating an outline of a protein structure P.

(ii) Compute a sequence of line segments from the sequence of lines obtained in step (i).

First we consider step (i). Let $LS = (L_1, L_2, \cdots, L_K)$ be a sequence of lines, and $I = (i_1, \cdots, i_{K+1})$ be a sequence of integer numbers such that $i_1 = 1$, $i_{K+1} = n$ and $i_k < i_{k+1}$. We define the score $FIT(P, LS, I)$ by

$$FIT(P, LS, I) = \sum_{j=i_1}^{i_2} d(\boldsymbol{p}_j, L_1)^2 + \sum_{j=i_2}^{i_3} d(\boldsymbol{p}_j, L_2)^2 + \cdots + \sum_{j=i_K}^{i_{K+1}} d(\boldsymbol{p}_j, L_K)^2 ,$$

where $d(\boldsymbol{p}_j, L_k)$ denotes the distance between a point $\boldsymbol{p}_j$ and a line L_k. Given P and K, it is possible to compute the pair (LS, I) which minimizes $FIT(P, LS, I)$ in $O(Kn^2)$ time, using the least-squares fitting technique and the dynamic programming technique (note that this is a variant of the K-link path problem [1]). Moreover, given P and a positive real δ, it is possible to compute the pair (LS, I) minimizing K such that $FIT(P, LS, I) \leq \delta$ in $O(n^3)$ time. Thus, we can obtain a sequence of lines LS such that K is the minimum and $FIT(P, LS, I) \leq \delta$. The

Figure 1: An example of a sequence of segments computed from PDB data of pdb4hhb.

choice of δ is important because δ affects the quality of the obtained sequence. Currently, we use $\delta = 2.2\mathring{A}$.

Next we consider step (ii). Step (ii) is very simple. For each pair of lines (L_k, L_{k+1}) $(1 \leq k < K)$, we compute a point $s_k = \frac{q+r}{2}$, where $q \in L_k$ and $r \in L_{k+1}$ are the points such that $|\overline{qr}|$ is the minimum. Moreover, we compute a point $s_0 \in L_1$ (resp. $s_{K+1} \in L_K$) such that $|\overline{s_0 p_1}|$ (resp. $|\overline{s_{K+1} p_n}|$) is the minimum. Finally, we obtain a sequence of line segments $SS(P) = (\overline{s_0 s_1}, \overline{s_1 s_2}, \cdots, \overline{s_{K-1} s_K})$.

It is expected that $SS(P)$ is a good approximation of an outline shape of P. We have already implemented this method, and an example is shown in Fig. 1.

3 Application to the comparison of 3D structures

The above method can be applied to the comparison of 3D protein structures. Although several variants can be considered, we describe a simple one here.

¿From a sequence of segments $SS(P) = (s_1, \cdots, s_K)$, we construct a string $STR(SS(P))$ as follows, where each s_i denotes a line segment. Let $c(s_i)$ be the centroid of s_i. For segments s_i and s_j, $l_{i,j}$ denotes the length between $c(s_i)$ and $c(s_j)$, $\alpha_{i,j}$ denotes the angle between s_i and s_j, $\beta_{i,j}$ denotes the angle between $\overline{c(s_i)c(s_j)}$ and s_i, and $\gamma_{i,j}$ denotes the angle between $\overline{c(s_i)c(s_j)}$ and s_j (see Fig. 2). For each s_i such that $i \leq K - D$, let $STR(s_i)$ be

$$((l_{i,i+1}, \alpha_{i,i+1}, \beta_{i,i+1}, \gamma_{i,i+1}), (l_{i,i+2}, \alpha_{i,i+2}, \beta_{i,i+2}, \gamma_{i,i+2}), \cdots, (l_{i,i+D}, \alpha_{i,i+D}, \beta_{i,i+D}, \gamma_{i,i+D}))\ ,$$

where D is an appropriate constant. Then, $STR(SS(P))$ is obtained by concatenating $STR(s_1)$, $STR(s_2)$, $\cdots$, $STR(s_{K-D})$, where concatenation of $(t_1, \cdots, t_p)$ and $(u_1, \cdots, u_q)$ is $(t_1, \cdots, t_p, u_1, \cdots, u_q)$.

Next, we define a score between $(l_{i,j}, \alpha_{i,j}, \beta_{i,j}, \gamma_{i,j})$ and $(l_{i',j'}, \alpha_{i',j'}, \beta_{i',j'}, \gamma_{i',j'})$ by

$$C_1 - C_2|l_{i,j} - l_{i',j'}| - C_3|\alpha_{i,j} - \alpha_{i',j'}| - C_4|\beta_{i,j} - \beta_{i',j'}| - C_5|\gamma_{i,j} - \gamma_{i',j'}|\ ,$$

where $C_1, \cdots, C_5$ are appropriate constants. Then, for two protein structures P and Q, we compute an optimal alignment between $STR(SS(P))$ and $STR(SS(Q))$ by means of a conventional

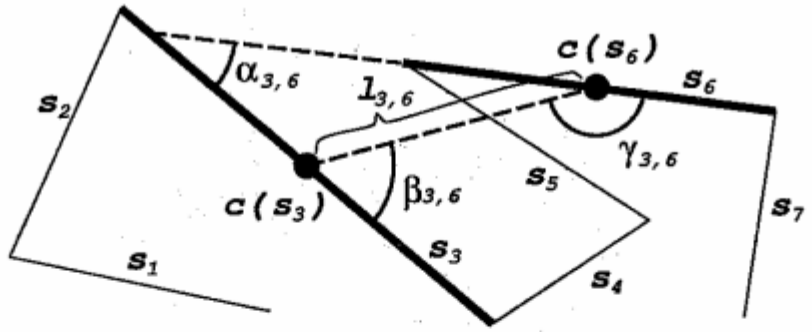

Figure 2: Definitions of $l_{i,j}$,$\alpha_{i,j}$,$\beta_{i,j}$ and $\gamma_{i,j}$ used in $STR(SS(P))$.

alignment algorithm for two strings, where each quadruplet $(l_{i,j}, \alpha_{i,j}, \beta_{i,j}, \gamma_{i,j})$ corresponds to a character. Finally, we consider the score of an optimal alignment as one indicating the similarity between P and Q. It is expected that the score is high if P is similar to Q. Note that not only local similarities but also global similarities are taken into account if large D is used.

Currently, we are examining this comparison method using PDB data. Details and results will be reported elsewhere.

4 Concluding remarks

We have proposed a new method for representing 3D protein structures. Although we have described one application only, we believe that it can be applied to other problems. For example, it might be applied to clustering of protein structures. Thus, applying the method to other problems is important future work.

The method might be modified for the case where line segments are replaced by special kinds of curves, and better fitting might be obtained. Thus, it is also important to study such variants.

References

[1] A. Agarwal, B. Schieber and T. Tokuyama, "Finding a minimum weight K-link path in graphs with Monge property and applications", *Proc. ACM Symp. Computational Geometry* (1993) 189–197.

[2] T. Akutsu, "Efficient and robust three-dimensional pattern matching algorithms using hashing and dynamic programming techniques", *Proc. 27th Hawaii International Conference on System Sciences* (1994) 225–234.

[3] R. Nussinov and H. J. Wolfson, "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques", *Proc. Natl. Acad. Sci. (USA)* **88** (1991) 10495-10499.

[4] W. R. Taylor and C. A. Orengo, "Protein structure alignment", *J. Molecular Biology* **208** (1989) 1–22.

[5] G. Vriend and C. Sander, "Detection of common three-dimensional substructures in proteins", *PROTEINS: Structure, Function, and Genetics* **11** (1991) 52–58.