PROTEIN STRUCTURE COMPARISON USING REPRESENTATION BY LINE SEGMENT SEQUENCES

TATSUYA AKUTSU HIROSHI TASHIMO
Department of Computer Science, Gunma University
1-5-1 Tenjin, Kiryu, Gunma 376, Japan [a]

1 Introduction

Classification of tertiary protein structures is very important for better understanding of protein structures.[3,4,5,6] In most of them, protein structure alignment is a basic tool for classification. Although structure alignment algorithms work very well in most cases, they may fail to find similar folding patterns in some cases because having similar folding patterns is not necessarily equivalent to existence of a good structure alignment. Moreover, most protein structure alignment algorithms require long CPU time.

On the other hand, classification by topology diagram has been used traditionally.[2] In this method, each protein structure is treated as a sequence of α-helices, β-strands and turns. Although this method is useful, it seems difficult to make the classification procedure automatic and difficult to classify structures into more detailed families.

Thus, we propose an intermediate comparison method. In this method, each tertiary structure is represented by a sequence of line segments and the similarity between two structures is measured by the score of the alignment between two sequences of line segments. Note that, once such representation (i.e., a sequence of line segments) is computed, the alignments can be computed quickly because most structures are represented by means of sequences of at most 100 line segments.

2 Method and Result

A sequence of line segments which approximates an outline of each input tertiary structure is computed using the least-squares fitting technique and the dynamic programming technique (see Fig. 1), where details are omitted here.

The comparison of two sequences of line segments is done via the following two steps. First, two input sequences of line segments are transformed into strings respectively, and then the score between two strings is computed applying the standard string (sequence) alignment algorithm (or the double

[a] **e-mail:** akutsu@cs.gunma-u.ac.jp tashimo@keim.cs.gunma-u.ac.jp

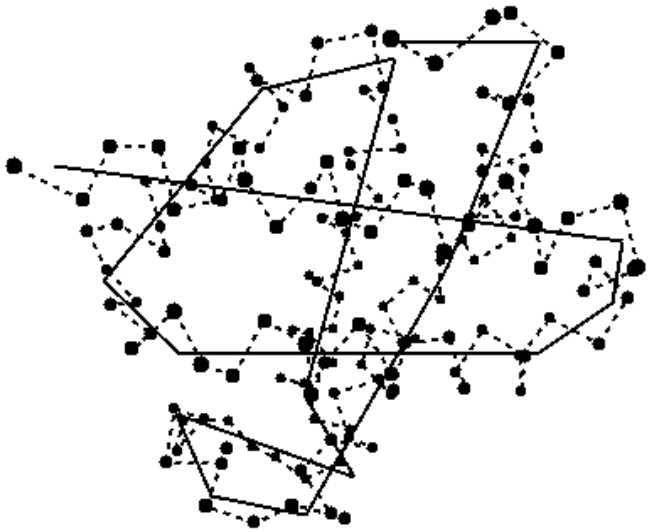

Figure 1: An example of a sequence of line segments computed from protein structure '4hhb'. Black circles denote the input Cα atoms and black lines denote the obtained line segments.

dynamic programming technique [7]). In the transformed strings, each character corresponds to a pair of line segments and the score between two characters is high if a pair of line segments is similar to the other pair of line segments.

The proposed method was compared with a structure alignment algorithm, which was previously developed by us,[1] using PDB data. The results show that the proposed method is much faster than the previous one and classifies tertiary structures at least as well as the previous one does.

References

1. T. Akutsu, *Proc. 27th Hawaii Int. Conf. on System Sciences* **5**, 225–234 (1994).
2. C. Branden and J. Tooze, *Introduction to Protein Structure*, (Garland Publishing, 1991).
3. L. Holm *et al.*, *Protein Science* **1**, 1691–1698 (1992).
4. C. A. Orengo *et al.*, *Nature* **372**, 631–634 (1994).
5. S. Pascarella and P. Argos, *Protein Engineering* **5**, 121–137 (1992).
6. A. Šali and J. P. Overington, *Protein Science* **3**, 1582–1596 (1994).
7. W. R. Taylor and C. A. Orengo, *J. Molecular Biology* **208**, 1–22 (1989).