Protein Threading Using a Score Function Derived by a Linear Programming Based Method

Tatsuya Akutsu [1] Hiroshi Tashimo [2]

takutsu@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

[2] Gunma University, 1-5-1 Tenjin, Kiryu, Gunma 376, Japan

Abstract

We have been developing a novel method of deriving a score function for protein threading. In this method, the constraint that the score of the native threading is minimum over all possible threadings is expressed in a form of linear inequalities, and then parameters defining a score function are determined by solving these inequalities. The proposed method was evaluated using Lathrop and Smith's algorithm for finding optimal threadings and was shown to be effective for computing nearly correct threadings.

1 Introduction

Protein threading is a method for the protein 3D structure prediction. In protein threading, an *alignment* between spatial positions in a structural model and amino acids in a sequence is computed using a suitable *score function* (equivalently, *potential function*). Precisely, an alignment which minimizes the total score (corresponding to potential energy) is computed, and such an alignment is called an *optimal threading*.

To use protein threading in a predictive setting, there are two major problems: *how to compute an optimal threading*, and *how to derive a score function*. For the former problem, although a lot of methods have been proposed, only the method by Lathrop and Smith can compute an optimal threading in reasonable CPU time [3]. For the latter problem, a lot of methods have also been proposed. In most of them, score functions are derived using Boltzmann-like statistical methods. However, using such a score function, it is not guaranteed even for training data that optimal threadings always coincide with *native threadings* (i.e., native structures). Therefore, we developed a new method for deriving a score function. Although Maiorov and Crippen have already proposed a similar method [4], they did not consider the effect of variable length gaps between core segments.

Since this is a short abstract, interested readers may refer Ref. [1] for details. Theoretical studies and extensions of the method for other problems are described in Ref. [2] too.

2 Method and Results

In the proposed method, a score function is expressed by a set of parameters. Then, from training data (known 3D structures), we randomly generate alternative threadings and make a constraint:

$$score(\text{native threading}) + gap \ < \ score(\text{alternative threading})$$

[2]Present address: Toyobo Co. Ltd., Tsuruga-city, Fukui 914, Japan.

for each alternative threading, where *gap* is an appropriate constant. Finally, we derive a score function by computing the values of parameters for which these inequalities are simultaneously satisfied. Since each constraint is *linear* (if we assume a usual contact potential), we can use a *linear programming* (LP, in short) solver. Note that, if we can generate all possible threadings, it is guaranteed for training data that optimal threadings coincide with native threadings. However, generating all possible threadings takes too long time and thus we use randomly generated threadings.

In order to evaluate the quality of the score function, we derived a score function from 24 protein data (training data) and computed self-threadings for 23 protein data (test data), where all data were non-homologous. These protein data sets are subsets of one used by Lathrop and Smith [3]. As in Ref. [3], we measured the displacement between the optimal and the native threadings for each core segment across all self-threading trials. Error distributions from 89 α-helix and 77 β-strand core segment threadings are shown in Fig. 1. Although our test data set is a subset, our results are as good as those in Ref. [3] where results across 5 score functions are shown. From these results, we can see that our score function is at least as good as existing score functions. Moreover, the proposed method has an advantage over statistical methods because our score function is derived only from 24 protein data, while most score functions are derived from hundreds of protein data.

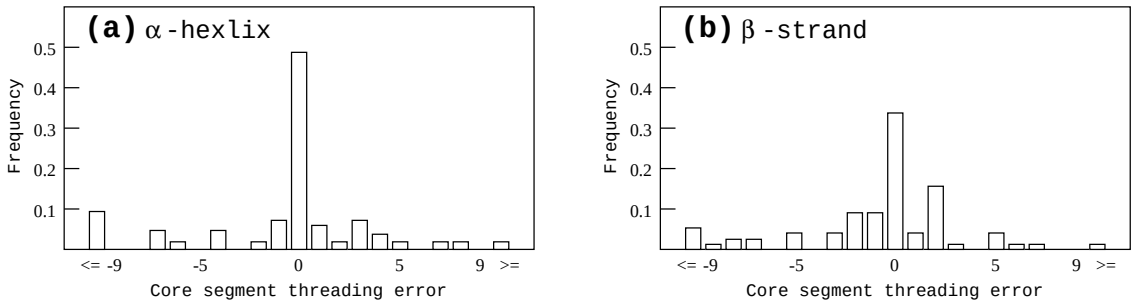

Figure 1: Frequencies of alignment errors between the optimal and the native threadings for each core segment across 23 protein data (89 α-helices and 77 β-strands). Error is computed as the optimal threading sequence index minus the native threading sequence index.

Acknowledgement

This work was supported in part by a Grant-in-Aid "Genome Science" for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Akutsu, T. and Tashimo, H., "Linear programming based approach to the derivation of a contact potential for protein threading," *Proc. Pacific Symposium on Biocomputing'98* (in press).

[2] Akutsu, T. and Yagiura, M., "Linear programming based approach for learning score functions in molecular biology," *Proc. Japan-Korea Joint Workshop on Algorithms and Computation*, 144–151, 1997.

[3] Lathrop, R. H. and Smith, T. F., "Global optimum protein threading with gapped alignment and empirical pair score functions," *J. Mol. Biol.*, 255:641–665, 1996.

[4] Maiorov, V. N. and Crippen, G. M., "Contact potential that recognizes the correct folding of globular proteins," *J. Mol. Biol.*, 227:876–888, 1992.