Accuracy of multiple sequence alignments as assessed by reference to structural alignments

O. Gotoh
gotoh@saitama-cc.go.jp

Department of Biochemistry, Saitama Cancer Center Research Institute
818 Komuro, Ina-machi, Saitama 362 Japan

1 Introduction

In the last 20 years, many multiple-sequence alignment programs based on various principles have been developed. Continuous efforts have been devoted to solve two major problems: (1) how to evaluate the 'goodness' of an alignment, and (2) how to get the alignment with the optimal score. These problems are tightly interrelated, and other criteria are needed to objectively assess reliability of a certain alignment method. Recently, the number of protein three-dimensional (3D) structures determined by X-ray crystallography and high-resolution NMR methods is rapidly increasing. Comparison of the 3D structures makes it possible to align distantly related protein sequences based on their structural equivalence. A few collections of such structure-based alignments are now available [4]. Hence we can assess the quality of sequence alignments obtained by a given method by referring to the structural counterparts. McClure et al. [3] recently reported that the! most popular 'progressive' metho

2 Methods

The three multiple sequence alignment strategies examined are (1) progressive method, (2) randomized iterative method without weight [1], and (3) randomized iterative method with weight [2]. In addition, the conventional pairwise sequence alignment method based on dynamic programming is also used as a control. A simple scoring system of Dayhoff-type PAM250 matrix was used as a measure of similarity between aligned amino acids. Except for the pairwise method, all members in a protein family are aligned, and the subalignment composed of the sequences of known structures are extracted from the larger set and compared with the structural

後藤　修：埼玉県立がんセンター研究所生化学部、〒 362 埼玉県北足立郡伊奈町小室 818

alignment. The degree of consistency of two alignments was evaluated as described previously [2]. Various combinations of gap opening and extension penalty values were examined, and the result with the best consistency was used for further analysis.

3 Results

The Joy3.2 database [4] of structural alignments consists of 110 entries. Because the present purpose is to assess accuracy of sequence alignment of distantly related members, those entries consisting of only two members, of closely related members (amino acid identities > 40 %), or of short sequences (length < 60 aa) are omitted, and total of 34 families were examined. The general tendency of the accuracy obtained by the four methods was as follows: pairwise alignment between single members < progressive method < unweighted randomized iterative method < weighted randomized iterative method. This tendency is exactly the same as that previously observed for the globin family [2], although the degree of consistency between sequence and structural alignments widely varied among the families examined. The above tendency is most prominent for large families consisting of divergent members, such as globins and cytochrome P450s. In these cases, th! e weighted randomized strategy pr

Acknowledgement

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

References

[1] O. Gotoh "Further improvement in methods of group-to-group sequence alignment with generalized profile operations," *Comput. Applic. Biosci.* vol. 10, pp. 379-387, 1994.

[2] O. Gotoh "A weighting system and algorithm for aligning many phylogenetically related sequences," *Comput. Applic. Biosci.* in press.

[3] M.A. McClure, T.K. Vasi, and W.M. Fitch, "Comparative analysis of multiple protein-sequence alignment methods," *Mol. Biol. Evol.*, vol. 11, pp. 571-592, 1994.

[4] A. Sali and J.P. Overington, "Derivation of rules for comparative protein modeling from a database of protein structure alignments," *Protein Sci.*, vol. 3, pp. 1582-1596, 1994.