Multiple Sequence Alignment with a Tree-Based Weighting System

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

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

Multiple sequence alignment is a powerful tool for studying structure-function relationship and evolution of biological macromolecules. The problem is computationally hard, and it is impractical to get an exact solution for more than several sequences. Although the so called progressive method is most widely used today for aligning a large set of sequences, the quality of alignments thus obtained is not satisfactory especially when the sequences to be aligned are distantly related. Recent progress in methods has changed the situation, and we can now obtain a high-quality alignment of a large number of sequences in practical computation time. The strategy attempts to refine a crude alignment by iteration with rigorous pairwise alignment between randomly partitioned submembers of the sequences [1, 2]. The cost of calculation for large sets of sequences can be suppressed by hundreds times by the use of 'generalized profile operations' [3].

Although the goodness of an alignment has been evaluated with the sum-of-pairs (SP) measure in these method, this measure is not appropriate when evolutionary distances between members are not evenly distributed. When we look at the globin superfamliy, for example, there are many alpha-chain and beta-chain haemoglobins and myoglobins with similar sequences within each family, while non-vertebrate globins are highly diverged from one another. Thus, the contributions of these minor members to the total SP score are predominated by those of large families. To correct the biased contributions, Altschul et al. suggested to calculate an SP score with a weight given to each pair of members, and proposed two methods for obtaining such a set of weights [4]. The weighted sum-of-pairs (WSP) score invokes a new problem, however, since profile based operations can no longer be applicable in general, which invalidates the accelerated alignment algorithm [3]. I show here that the profile-based fast algorithm of multiple sequence alignment can be regained under some limited yet practical conditions imposed on the weights.

In brief, we first calculate an unrooted tree which represents evolutionary relationships among all the members of sequences to be aligned. We impose a limitation on the weight

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

given to a pair of members, W_{ab}, so that it is expressed as a product of factors assigned to the edges along the path from the member **a** to **b** in the tree. One of the Altschul et al.'s weighting system [4] intrinsically satisfies the above limitation. Although the other, more sound weighting system does not meet the condition *per se*, we can easily obtain a very good approximation in accordance with the limitation. When the tree is dissected into two groups at an edge, the weight given to a pair between one member of the left side of the tree and that of the other side is 'factorized', and so it is possible to define the profiles of the left- and right-side groups. In this way we get the WSP of a given alignment in $O(N)$ operations in stead of $O(N^2)$ with the conventional method, where N is the number of primary sequences in the alignment. Refinement of a crude alignment by the randomized iteration method is also carried out without a significant loss of efficiency if partition points are restricted to the edges in the tree. Comparing the results of sequence alignment with those derived from three- dimensional structure of proteins, we found a general tendency in the quality of alignments obtained with various sequence alignment methods, *i.e.* the matching of the sequence alignments to structural ones improves in the order of the used methods: (1) pairwise alignment between single members $<$ (2) progressive method $<$ (3) randomized iteration without weight $<$ (4) that with weight.

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 of Japan.

References

[1] M.P. Berger and P.J. Munson, "A Novel randomized iterative strategy for aligning multiple protein sequences," *Comput. Applic. Biosci.*, Vol. 7, pp. 479-484, 1991.

[2] O. Gotoh, "Optimal alignment between groups of sequences and its application to multiple sequence alignment," *Comput. Applic. Biosci.*, Vol. 9, pp. 361-370, 1993.

[3] 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.

[4] S.F. Altschul, R.J. Carroll, and D.J. Lipman, "Weights for data related by a tree," *J. Mol. Biol.*, Vol. 207, pp. 647-653, 1989.