An Accurate Multiple Sequence Alignment Algorithm for Biological Sequence Sets with High Length Variations

J.A.D.T.B. Jayasingha*, C.T. Wannige

Department of Computer Science, University of Ruhuna, Matara, Srilanka.

Multiple sequence alignment (MSA) is used for many studies in modern biology. There are many algorithms available for the alignment of multiple sequences. Among them, progressive alignment algorithm is the most commonly used heuristic alignment strategy for MSA. It solves MSA with an economic complexity but does not provide accurate solutions, because there is a conflict between accuracy and complexity. The existing similarity score method in progressive alignment algorithm does not consider the lengths of the sequences in the considered sequence set. So, if the protein or DNA sequences are with high length variations, the initial alignment scores may not produce a correct measure of similarity between the sequences. This leads to less accurate initial alignment scores, and as a consequence, final multiple sequence alignment may produce inaccurate results. In this research, we present a modified progressive alignment algorithm especially for sequences with high length variations. We modify the latest version of ClustalW 2.1 by replacing the similarity distance measure in ClustalW algorithm with a novel distance measure. The new distance score method captures the distance between each sequence pairs in sequence set and the obtained distance measure is utilized to generate a better guide tree for progressive alignment. In order to determine the pairwise similarity distance measure, we used lengths of the *shortest common super-sequence (SCS)* and the *Longest Common Sub-sequence (LCS)*. We assessed our algorithm with BALIBASE 3.0 protein benchmark and compared the obtained results to those obtained with ClustalW alignment algorithm using the Quality score (Q Score) and the Sum of Pairs Score (SPS). We obtained better Q scores and SP scores for the alignments from modified ClustalW algorithm over original ClustalW algorithm. Furthermore, the alignment speed of modified ClustalW algorithm is multiple times faster than the original ClustalW algorithm.

Keywords: Multiple sequence alignment, Distance measure, Shortest Common Super-sequence, Longest Common Sub-sequence.