Eulerian Path Methods for Multiple Sequence Alignment

Michael S.Waterman
Department of Biological Science
University of Southern California
Los Angeles, CA 90089-1340
msw@usc.edu

Yu Zhang
Department of Mathematics
University of Southern California
Los Angeles, CA 90089-1113
yuzhang@usc.edu

Abstract

With the rapid increase in the size of genome sequence databases, the multiple sequence alignment problem is increasingly important and often requires the alignment of a large number of sequences. Beginning in 1975, many heuristic algorithms have been created to improve the speed of computation and the quality of alignment. We introduce a novel approach that is fundamentally distinct from all currently available methods. Our motivation comes from the Eulerian method for fragment assembly in DNA sequence determination, that transforms all the DNA sequencing fragments into a de Bruijn graph and then reduces sequence assembly to a Eulerian path problem.

This lecture focuses on global multiple alignment of DNA sequences, where entire sequences are aligned into one configuration. The main result is an algorithm with almost linear computational speed with respect to the total size (number of letters) of sequences to be aligned. In a simulation, 500 sequences (averaging 500 bases per sequence and as low as 70% pairwise identity) have been aligned within 3 minutes on a personal computer while the quality of alignment is satisfactory. As a result, accurately and simultaneously aligning thousands of long sequences within a reasonable amount of time becomes possible. Data from an Arabidopsis sequencing project is used to demonstrate the performance.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**