SNAP: A Novel Algorithm for Fast Global Sequence Alignment and Database Search

Sivaraman, Venkatesh

There has been an explosion in the amount of genetic data available to biologists in the past two decades. To analyze this vast information, computer algorithms (e.g. BLAST) have been developed that perform sequence alignment to compare many genetic sequences. However, considering the sheer magnitude of data (170 million sequences in GenBank alone), the goal of this project was to create an algorithm that aligns sequences faster than previous algorithms while maintaining accuracy. The new algorithm, called Segmented Numerical Alignment Preprocessing (SNAP), uses an approximated scoring matrix and a grouping technique to reduce the number of steps needed to align two sequences. The amino acid sequences are divided into small groups and converted into numbers to be aligned. Therefore, SNAP aligns sequences at a lower “resolution” than other algorithms, but at a much higher speed. SNAP was tested alongside four other algorithms: Needleman-Wunsch (the gold standard for accuracy), simulated annealing, a genetic machine learning algorithm, and BLAST. For two test datasets, one containing 189 sequences and the other with 6,724 sequences, ten query sequences were processed by each algorithm to find the highest-scoring alignments in the dataset. In terms of accuracy, SNAP was superior to simulated annealing and the genetic algorithm and comparable to BLAST. It also performed on average 70 times faster than BLAST. Based on these findings, it is proposed that SNAP is a viable alternative to BLAST for searching today’s large sequence databases.

Awards Won:

Third Award of $1,000