Abstract

Nahla A Belal

A Theoretical Model for Whole Genome Alignment

We present a graph-based model for representing two aligned genomic sequences. An alignment graph is a mixed graph consisting of two sets of vertices, each representing one of the input sequences, and three sets of edges. These edges allow the model to represent a number of evolutionary events. This model is used to perform sequence alignment at the level of nucleotides. We define a scoring function for alignment graphs. We show that minimizing the score is NP-complete. However, we present a dynamic programming algorithm that solves the minimization problem optimally for a certain class of alignments, called breakable arrangements. Algorithms for analyzing breakable arrangements are presented. We also present a greedy algorithm that is capable of representing reversals. We present a dynamic programming algorithm that optimally aligns two genomic sequences, when one of the input sequences is a breakable arrangement of the other. Comparing what we define as breakable arrangements to alignments generated by other algorithms, it is seen that many already aligned genomes fall into the category of being breakable. Moreover, the greedy algorithm is shown to represent reversals, besides rearrangements, mutations, and other evolutionary events.