12/11/2021

Mehmet Demirci
182010020018
Şükrü Berke Çakır
172010020025

CENG4217 Bioinformatics Algorithms Project Proposal

Problem

All DNA sequencing machines have a limited read length and an error profile. Each sequence read from DNA sequencing machine is called a **Read**. We need to combine these reads and construct DNA sequence. We have several algorithmic ways to combine these reads to construct a DNA sequence. While current methods produce very good results, the perfect assembler has yet to be built.

With given algorithms, which algorithm is best suited for the given dataset. That is our question. To understand the differences my partner and I collaborated with another group of 2 and decided to split two algorithms with same dataset and run tests over that dataset with given algorithm and compare results (i.e., run time, cost, complexity, accuracy, etc.)

Algorithms

My partner and I decided to implement de Bruijn Graph Assembly algorithm to assemble given dataset. And the other group decided to implement Overlap Layout Graph Assembly algorithm to assemble same dataset.

1. de Bruijn Graph Assembly

This algorithm works such that: For a given small integer k, consider every k-mer (i.e., substring of length k) of every read, and consider two k-mers a and b to overlap if the last $k-1$ symbols a exactly match the first $k-1$ symbols of b. This k-mer "overlap graph" is the de Bruijn graph of order k. To construct a genome path using de Bruijn graph is via a method called **Eulerian path**.

2. Overlap Layout Graph Assembly

This algorithm works such that: For a given small integer k, consider ever k-mer(i.e., substring of length k) of every read as vertices, and connect every vertex a to b if suffix of length k-1 of a is equal to prefix of length k -1 of b via an edge. This constructed graph is the Overlap Layout graph of order k. To construct a genome path using Overlap Layout graph is via a method called **Hamiltonian path**.

3. Eulerian Path vs Hamiltonian Path

- An **Euler path** is a path that passes through every edge exactly once. If it ends at the initial vertex then it is an Euler cycle.
- A **Hamiltonian path** is a path that passes through every vertex exactly once (NOT every edge). If it ends at the initial vertex then it is a Hamiltonian cycle.