Project description – 21/2/2018

Title: Lightweight streaming algorithm for sequence assembly

Background: Modern DNA sequencing methods produce hundreds of millions of short sequences called reads, each of length 50-200 letters. The locations of the reads in the genome are unknown. The first basic step in the study of new genomes and in comparing between genomes is assembly: solving a huge puzzle of combining the short reads into long sequences, ideally representing whole chromosomes. Many modern assemblers were developed to date, but they are extremely resource intensive, using a lot of CPU time and memory. Moreover, in some cases even storing the raw read data locally on the user's machine is impossible or too expensive. In that case a streaming solution is needed: downloading the read from a remote site and forming an assembly without storing the reads locally. We have recently developed one of the fastest and most resource efficient assemblers, called Faucet, which is currently the best streaming assembler. It assembles the reads based on two streaming rounds of the read data.

Goal: Develop and implement a single-round streaming algorithm on top of Faucet, improve Faucet, and conduct experiments to measure performance of the new algorithm.

Details: Currently, Faucet streams through all of the sequence reads twice in order to process them. In the project, you will (1) modify the algorithm to process reads in a single streaming step; (2) implement the algorithm as a multi-threaded parallel program; (3) design and run experiments to test the new algorithm on real sequencing data; and (4) develop processing methods that improve assembly quality. If successful, we intend to publish a paper describing the algorithm and its performance.

Prerequisites: C/C++ programming experience. High motivation to work on sequence assembly, data structures, and algorithms

Project leader: David Pellow

Reading material: Rozov, R, Goldshlager, G, Halperin, E, Shamir, R. (2017). Faucet: streaming de novo assembly graph construction. Bioinformatics, 34(1), 147-154.

Code base: https://github.com/Shamir-Lab/Faucet