Computational Biology Independent Project
Prof. Anna Ritz

Eliotte Garling
Lab Date: 05/02/17

Problem Statement

The development of Next-Generation sequencing (NGS) technologies has proceeded at an unprecedented pace, leading to two major bottlenecks in determining the best way to assemble and extrapolate information from genomes. The first has to do with the genome assembly process, the second is annotation. Both processes are significantly hindered by repetitive sequences, which make up large portions of most genomes. The main motivation of the project is to write a program which will filter sequences into low, medium and high complexity sequences – ideally speeding up the process of genome annotation.

Data Formatting

During the initial stages of the project I used a toy sequence from Homework 5, then implemented a 4090 bp portion of a *Daphnia magna* mitochondrial genome that I previously assembled.

Steps of the Program

The first part of this project is adapted from the frequent words problem of homework 5. A sequence is divided up into *k*-mers then the most and least frequent are separated. The most

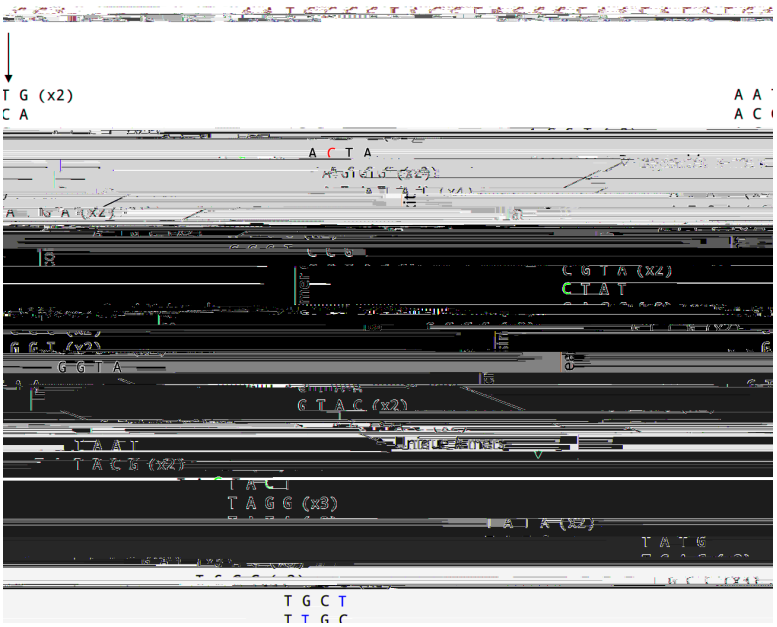

Figure 1: Diagram of the project steps used to organize genomic sequences into k-mers of relative complexity and frequency to improve the genome annotation problem. (Portions of figure adapted from Wikipedia Commons)

frequent presumably, represent retrotransposons or transposable elements (TEs), which are capable of replicating in the genome and have highly conserved sequences. The 'uncommon *k*-mers', or *k*-mers, which don't have identical sequences, are put into a frequency table to generate a consensus sequence. The hamming distance of the infrequent *k*-mers to the consensus sequence is measured, and *k*-mers are grouped as highly complex if they had little similarity with the consensus sequence, and semi-complex if they had higher similarity with the consensus sequence. The level of similarity can be adjusted, so that the threshold for uniqueness can be determined by the user by selecting a proportion between 1 and 0. Ideally, this program would allow quick access to high complexity sequences, since they should be functional sequences and give a decent idea of the unique functional aspects of the genome. This project was inspired by a paper by Anvar et al. 2014, which described the formation new open-source database called kPAL, which allows for the rapid characterization of functional and repetitive sequences in a genome by analyzing *k*-mer sequences.

Discussion

Overall the program was successful in it's categorization of *k*-mers based on the relative frequency, and complexity compared to a consensus sequence. I discovered that automatically formating to a FASTA or FASTQ format would be ideal so that multiple *k*-mers could be BLASTed at once. However, upon BLASTing a 200 bp k-mer length where sequences with more than 50% similarity to the consensus sequence were ltered into the 'semi-unique *k*-mer' category. And sequences with less than 50% similarity to the consensus sequence were ltered into the 'Unique *k*-mer' category. A BLAST search of a *k*-mer from the repeated *k*-mer category, was a 97% identity, and had

Biblography

Anvar, S. Y., Khachatryan, L., Vermaat, M., van Galen, M., Pulyakhina, I., Ariyurek, Y., Laros, J. F. (2014). Determining the quality and complexity of next-generation sequencing data without a reference genome. Genome Biology, 15, 555. https://doi.org/10.1186/s13059-014-0555-3