Project 1c,1d: Metagenomics
Project Description version 1.

The goal of this project is to perform metagenomic analysis on a sample.

The input to the project are many genome sequences which may be in a sample and reads from the sample. The output is a prediction of the source genome for each read.

A certain number of genomes are present in each sample. Reads are generated the same way as in Project 1a. There are some repeated regions in each genome.

The “answer” file that is uploaded to Codalab is a “read header with sources” file. This file contains the headers of the reads along with the source genome of the read. The predicted sources are compared to the true sources and the score is based on how many reads are classified correctly. There are only a small number of genomes in the sample. However, there are duplicated regions between genomes so a read may map to more than one genome. The correct “source” is most likely the more common sources.

You can use your solution of Project 1a,b to solve this problem by mapping the reads to each genome. This should be enough for Project 1c. For Project 1d you may need to use some of the advanced data structures (Bloom Filters, Minimizers, Minimum Perfect Hashing, etc) to get it to scale to the larger data.

A sample metagenomics problem is available to help understand the format.

The files provided for the sample metagenomics problem are:
project1c_sample_genome_N.fasta - The reference genome for the “N”s organism. If this organism is in the sample, then you will see reads from this organism in the sample. There will be many of these.
project1c_sample_reads_with_source_and_positions.fasta - This file contains the reads with the genome source and position listed in the header. This is here just for the sample problem so you can double check your work.
project1c_sample_reads.fasta - This is the reads file without the source and position information in the headers. This is the input for the project.
project1c_sample_answers.txt - This is the answers file for the sample metagenomics problem.

The files provided for project 4a are:
project1c_genome_N.fasta - The reference genome for the “N”s organism. If this organism is in the sample, then you will see reads from this organism in the sample. There will be many of these.
project1c_reads.fasta - This is the reads file that will be used to predict the organisms present in the sample.

The files provided for project 1d are:
project1d_genome_N.fasta - The reference genome for the “N”s organism. If this organism is in the sample, then you will see reads from this organism in the sample. There will be many of these.
project1d_reads.fasta - This is the reads file that will be used to predict the organisms present in the sample.

General Hints:
There are many ways to solve this problem so feel free to try anything.
One straightforward way is to use your solution to project 1a and build an index for each genome. Then map your reads to each genome one at a time. You do not need to store all of the indices for each genome in

memory. You can do this one at a time. Once you get the mapping for each of the genomes, combine this file to generate the final file.

Other solutions are much faster by building an index for all of the genomes at the same time using the techniques in class.

Note that there are a small number of genomes in each sample. Each genome has some duplicated sequence of other genomes. So reads will map to more than one genome. However, you can use the number of reads that map to each genome to figure out which genomes are actually present and which genomes just have copies of part of a present genome. You can use a threshold to figure out which ones are present.

Once you figure out the small set of genomes that are present, you can then run your solution on just those genomes which will assign each read to only one of them.

Good luck!