Milestone 1 Report

Sunny Nahar **12/10/15**

Major Changes:
There have not been any major changes in the goal of this research. The goal remains to find heuristics to optimize seed frequency in read mapping.

What You Have Accomplished So Far:
I have read the relevant literature regarding seed-and-extend based mappers, such as FastHASH, Hobbes, etc. This gave me familiarity with the architecture of seed-and-extend based mappers and where and how exactly seed frequency optimization is required. I implemented (almost) all of the testing infrastructure. This includes storing and loading the reference, loading seeds and testing a frequency predictor on a set of seeds, generating random sets of seeds, etc. This part is complete as of now, but I may need to add additional components as I see fit as the research goes forward. I implemented the basic frequency prediction models, uniform and threshold, to test the functionality of the infrastructure. I also implemented Cheap K-mer selection and Hobbes seed selectors. I started working on implementing the first of my ideas, a bidirectional frequency predictor. I anticipate completing this soon.

Meeting Your Milestone:
I have met the milestone I initially set. I read all the relevant literature, came up with a few ideas for optimizing seed selection frequency, and implemented the test infrastructure. These were the initial goals.

Surprises:
A surprise, which was not entirely unexpected, but the magnitude was unexpected, was the amount of time testing takes. Reading in the genome from disk takes about 1-2hrs, which means each test, even if it is a minor change, takes 1-2hrs. I initially got around this by running the tests on smaller (cut) pieces of the genome, but for accuracy tests, the entire genome must be used. I am trying to parallelize some of the code using OpenMP to speedup this process so that testing becomes smoother.

Revisions to your 15-400 Milestones:
I implemented the uniform and threshold seed predictor, which were the goals of the first milestone for next semester, so this is already done. I also implemented the Hobbes and Cheap Kmer selection predictor which were the goals of the second milestone, so this is done as well. I started implementing one of the heuristics, and did some preliminary testing, so this is the substance of some of the third milestone.

The time requirement to correctly implement the seed predictors was much larger than I expected. I anticipate I will need a lot more time to implement and develop each of the seed predictors, so I am extending the time period on the 3rd and 4th milestone by 2 weeks each. This comes from the 2 weeks from the 1st and 2nd milestone which is already done.

Resources Needed:
I have access to all the resources I currently need, which is just the reference genome and a few debug genomes. I have been using randomly generated seed sets instead of actual seed tests for testing. I will be able to retrieve actual seed sets when I need them. I have access to the 1TB RAM machine for testing my code as well.