Final Project Proposal

Nathan Roach, Patricia Walchessen, Charlie Wang, and Yue Zhang

October 26, 2017

Research Topic

SimHash is a widely used Locality Sensing Hashing (LSH) algorithm used to determine cosine similarity in large data sets. In 2016, Ondov et al. used MinHash, another LSH algorithm, to analyze the resemblance between genomes and metagenomes. This suggests that SimHash could also be used to analyze the resemblance between genomes and metagenomes. We would like to implement the SimHash algorithm to determine genetic similarity.

Implementation of SimHash would be a slightly different approach as MinHash analyzes resemblance whereas SimHash analyzes cosine similarity, but could prove scientifically useful as cosine similarity is more readily understood in scientific circles. In a paper comparing the results of MinHash vs. SimHash, MinHash outperformed SimHash when data was binary in areas of low and high similarity. Ideally, we will run a comparison of MinHash and SimHash on the same genetic data-sets and determine the optimal algorithm to run for genome similarity.

To evaluate our results, we will use the same criteria as Ondov et al. to maintain standards.

Input Data

1. S. cerevisiae
2. S. pombe
3. E. coli

Milestones

1. Replicating MinHash Data
2. Implementing SimHash algorithm
3. Benchmarking MinHash vs SimHash

Stretch Goals

1. Expand to full genome or multiple full genomes with varying GC content, redundancy etc.
2. Determine optimal k