# Introduction

Previously the methods used to compare metagenomes relied heavily on brute force methods and search algorithms that resulted in very inefficient algorithms not making it a viable method. However, my mentor recently proposed a method that would work at the gene level and would involve clustering genes for enhanced comparison. To improve upon this method, I will replicate this process on the protein level with a more efficient method in order to improve the overall performance.

The primary method of improving the clustering method is to not work with the original sequences but work with compressed versions of DNA sequences. Therefore it is possible to use a method that is very high in accuracy but preforms very poorly in terms of speed, and improve its speed by using compression. If compressed well enough while still retaining the ability to compare compressed sequences, the process can greatly be improved.

In this experiment, I tested 6 possible compression methods and was able to create an algorithm that was based on a byte array compression algorithm (Google’s Lz4 compression algorithm) and an appropriate similarity compression method using a compression distance measurement (CDM). This algorithm performed comparable in speed to the fastest algorithms currently in the industry while retaining the accuracy of the slower algorithms.

The efficiency of comparing metagenomes is a key part of understanding complicated diseases like cancer. Comparing the genomes of a healthy person and creating an accurate comparison with a diseased person can provide insight on possible factors that are different as a result of the disease. This in turn can be used as flags for early diagnosis as well as input towards a cure.