**Creating an efficient compression based DNA sequence clustering algorithm for the analysis and comparison of metagenomes**

Executive Summary

Metagenomics is the sequencing and analysis of genetic material from environmental samples. Metagenomics seeks to assess the content and diversity of microorganisms living in these samples. Unlike traditional genomics, Metagenomics is faced with the difficulty of encountering new mutations and varieties of microorganisms never before cultivated. As a result, DNA clustering methods were created to sort the sequences by similarity in order to analyze these clusters rather than individual sequences. However these methods have a long way to go in order to make Metagenomics accessible and accurate enough for common use. My compression based approach bridges the gaps in current clustering methods and improves the efficiency and accuracy. By first compressing the sequences, we work with smaller parts of data allowing the increase in efficiency. The approach describes a method allowing the combination of any compression method with any currently existing clustering algorithm. This approach was rigorously tested with three common compression algorithms, and a basic clustering algorithm on a wide variety of environmental data samples. We concluded that this approach lead to algorithms with efficiency, diversity and accuracy superior to the current leading clustering algorithms. For future research, we plan to use this approach on a state-of-the-art compression algorithm in conjunction with a complex clustering algorithm in order to create a clustering algorithm capable of increasing efficiency to the point that Metagenomics becomes an accessible lab practice. The efficiency of comparing metagenomes is a key part of understanding the role that microorganisms play in everyday situations. By comparing the environment inside a healthy and diseased person’s gut, we can analyze the importance and role of microorganisms in diseases. In order to better understand the role microorganisms play in our life, Metagenomics needs to become a more viable method of observation for complex samples.