Ashwin Sekar

SRP A

# Materials and Methods

## Materials

### Software

|  |  |  |
| --- | --- | --- |
| Software | Details | Developer |
| Java JDK 7 | Development Kit for Java on a Windows Environment | Oracle |
| Eclipse Juno | IDE for Java Development | Eclipse Foundation |
| Windows 8 Enterprise | Windows Operating System tailored toward large Buisnesses | Microsoft |

### Datasets

|  |  |
| --- | --- |
| Name | Details |
| UBA Acid Mine Drainage Biofilm (4441138.3) | <http://metagenomics.anl.gov/metagenomics.cgi?page=DownloadMetagenome&metagenome=4441138.3> |
| Data set from North Atlantic Deep Water and Axial Seamount, Juan de Fuca Ridge | DNA sequences used in Microbial diversity in the deep sea and the underexplored "rare biosphere". |

## Methods

### Evaluating compression based DNA similarity algorithms on ranked matches

#### Comparison by distance

The three compression based algorithms Sequitur, Lz4, and RLE were applied to the DNA sequences in the 2 data sets above. These compressed sequences were then compared to each other and then ranked based on most similar sequences.

#### Sequitur

This algorithm implementation involved taking a sequence and creating a set of grammar rules to represent the sequences. Separately, a database of these rules would be stored for reference. All the sequences operate on the same set of rules and are represented by which rules they use and the order of them. The rules must be used at least twice in order to be a valid rule and must contain at least 2 or more non terminal characters.

#### Lz4

Google’s Lz4 algorithm is a very fast lossless compression algorithm, providing compression speed at 300 MB/s per core, scalable with multi-cores CPU. It also features an extremely fast decoder, with speed in GB/s per core, typically reaching RAM speed limits on multi-core systems.

#### RLE

Run Length Encoding is a compression method which involves taking large repeating sequences and representing them with a counter on the amount of repetition.

These three algorithms were used to rank the sequences based on similarity to each other. Finally the original set of sequences was ranked using the current biologically sound method which involves edit distance and local sequences alignment. The rankings based on compression will be compared to the rankings on the original set and the amount of matches will be calculated.

### Comparison by CDM

CDM is a measure of similarity based on the compressed sizes. It involves concatenation of the 2 sequences and comparing the compressed size of the concatenation with the compressed size of the individual sequences. All three of the above algorithms were one again applied to the data set, this time using CDM as a measurement of similarity and ranking them. After they were ranked, the rankings were compared to both the edit distance and local sequence rankings of the original set.

### Further evaluation of the algorithms through clustering

After filtering out the inaccurate algorithms through the ranked matches’ evaluation, the promising algorithms are then evaluated through clustering.

A random DNA sequence within the data set is chosen as a cluster representative. Then all sequences with a high similarity measure to that representative are filed as part of that cluster. This process is repeated until all of the sequences have been clustered. The measure of similarity is chosen by the compression algorithm.

The results of the clusters are then compared with the biologically known results of the data. If done correctly, each cluster should correlate with a species or a sub species with little or no error in the sequences. Statistical parameters such as the CHAO1 index, Shannon index, and ACE index were calculated and compared to evaluate the results.

### Developing an alignment and universally applicable algorithm

Discrepancies in the clustering is sometimes a result of the randomness of the sequenced DNA strands. Sometimes the sequences are of uneven lengths and as a result cover more or less of the genes needed for clustering. A process by which the algorithm can find key proteins and use them as benchmarks to slice the sequences into smaller, even, and uniform chunks is to be developed so that all DNA sets can be compared regardless of preparation.