Background

With the recent advances in sequencing technologies, the amount of available DNA and RNA data is growing exponentially[1]. Such a high growth rate has made it necessary to remove the redundancies and reduce the storage space of the sequence data. In recent years, a common way of doing this is through clustering [?] [?], which can help to adress other biological questions [?] [?].

Clustering is defined as the grouping of uncategorised objects such that objects in the same group (cluster) have a higher similarity than those in other groups (clusters). A notable word here is uncategorised, which means that there is no prior knowledge about the objects. Habitually, the grouping is determined by a distance metric, so that objects closest to each other are placed in the same cluster. As a testament to the plentitude of ways there are of doing this is the number of clustering algorithms available [?]. Among these, `UCLUST` has repeatedly been shown to be the best clustering algorithm for RNA [?] [?]. Using a centroid based approach inspired by ICAtools [?], it has achieved speeds unheard of, without sacrificing much precision. Unfortunately, access to the 64-bit version of `UCLUST` requires a costly subscription.

The Molecular Microbiology Ecology Group of the Department of Biology, University of Copenhagen are amassing huge amounts of sequence data which they have no other tool than `UCLUST` to cluster. Searching for a free alternative to the costly algorithm, they sought out computer scientists to help them develop this free alternate.

In the course of this project, I have tried to develop an algorithm that could compete with `UCLUST`. Having confidence that Edgar, R.C., the creator of `UCLUST`, has optimized his approach efficiently, I used a different approach to the centroids based on. Studies have suggested that MinHash has approached the speed and precision of `UCLUST` [?]. Modifying this approach furthermore for increased speed, i hoped to achieve speed that could rival that of `UCLUST` without losing precision[2].

Problem Statement

I will attempt to create a clustering algorithm which can compete with, or even surpass, the speed of `UCLUST` without loss of precision, when working with huge datasets of sequences, each of lengths between 500 and 1500.

Delimitation

As we are mostly interested in the computational aspects of the algorithm, the biology behind the sequence data will not be described.

[1] `http://www.ncbi.nlm.nih.gov/genbank/statistics`

[2] Also, the Microbiology Group promised a good Scotch to the developer of an algorithm that could compete with `UCLUST`, which would fit right into the empty space on my shelf.

We only wish to test the speed of the algorithm on RNA and DNA, which is why files with the same format as fasta files are the only ones that will be accepted as input.

The version of the `UCLUST` i am comparing my algorithm with is **usearch8.0.1517_win32**, using the **cluster_fast** option. Therefore, i cannot guarantee that my algorithm will be better than newer versions of **UCLUST**.

Related Works

My main inspiration for this work was *Rasheed and Rangwala* 2013 [**?**], in which a MapReduce MinHash clustering algorithm was developed. What inspired be to improve upon this framework was *Jianqiu Ji et al.* 2013 [**?**], who the same year found a way to double the speed of MinHash without losing precision. Their approach was quite ingenious and simple, and with only a slight modification to it I developed the final method which turned out to be the seminal algorithm of this work.