1 Test Data

For the experiments, it was important that the data used was representative of real life data sets. Therefore, all data used were DNA and RNA sequences from bacteria and a multitude of Eukaryota. As we shall perform two different sets of experiments, each will have a seperate data set for tests; these will be described below.

1.1 Data used for precision tests

In order to test the precision of both algorithms, it was necessary to have relatively small sets, since we had to calculate a gold standard[1]. We used a clean sample of Cecum DNA data consisiting of 81,029 sequences. These were divided into 5 seperate sample files, each of which contained a sample of 10,000 sequences. These files will be refered to as `Cecum1`, `Cecum2`, `Cecum3`, `Cecum4`, and `Cecum5`. Each file contained a unique sample of the original file, none of which intersect with any of the other files.

1.2 Data used for speed tests

For the speed tests, two seperate files were taken in use, one DNA and one RNA. The DNA sample was of uncultured Actinobacteria, consisting of 2,879,170 sequences. This file was divided into five samples for the speed tests.

1. `Actino50K`: A sample of 50,000 sequences. Size = 72.5 MB
2. `Actino100K`: A sample of 100,000 sequences. Size = 146.6 MB
3. `Actine200K`: A sample of 200,000 sequences. Size = 283.5 MB
4. `Actino500K`: A sample of 500,000 sequences. Size = 663.4 MB
5. `Actino1mio`: A sample of 1,000,000 sequences. Size = 1286 MB

The SILVA SSU 119[2] database was taken into use as the RNA sample for the speed tests. It consists of all aligned sequences from both bacteria and a multitude of Eukaryota with a high alignment identity, without any sequences with 99% similarity. Just as the DNA sample, five samples from this file were used for testing

1. `Silva50K`: A sample of 50,000 sequences. Size = 80 MB
2. `Silva100K`: A sample of 100,000 sequences. Size = 160 MB
3. `Silva200K`: A sample of 200,000 sequences. Size = 320 MB
4. `Silva500K`: A sample of 500,000 sequences. Size = 790.5 MB
5. `Silva1mio`: A sample of 1,000,000 sequences. Size = 1573 MB

[1] which is a very costly affair

[2] `http://www.arb-silva.de/documentation/release-119/`

All samples consist of clean sequences, meaning they only contain the four characters DNA or RNA consist of. As one may note, the RNA samples are bigger than the DNA samples. This is caused by the fact that the sequences are longer in the RNA samples. These two files were chosen to see how the size of the sequences would affect the speed of the **MM** and $\mathbf{MM}\frac{1}{2}$ algorithms compared to `Uclust`.

1.3 Hardware

A Lenovo IdeaPad Y500 laptop was taken in use for the tests. Its processor is a Quad-Core 2.4 GHz, it has 8 GB RAM, and finally a hybrid HD with 25 GB SSD and 1 TB HDD. As only one computer was taken in use for the tests, the MapReduce Framework was not used to its full potential. With more computers, the speed of both the greedy algorithms for **MM** and $\mathbf{MM}\frac{1}{2}$ could be increased drastically.