Discussion

Looking back at the project, there are a few things that i would have done differently had i had more time.

First of all, the use of MapReduce had both pros and cons for my project. As i only had one computer to run the tests on, the distributed aspect of MapReduce was not taken in use. Higher speeds could therefore probably have been achieved with a more resource-optimized approach in a non-distributed environment. However, using another setup than MapReduce might have caused memory problems which MapReduce doesn't run into.

The results of the tests gave an exciting prospect for the **MM** algorithm. However i wish they had been more comprehensive, so that they could have cemented **MM** furthermore as a good contender. For the precision tests, I would have extended and done the following things:

- I would have used another precision measure. The error metric does not properly show how similar the clusters in the Gold standard were to those in **MM**,**MM**$\frac{1}{2}$ or `uClust`, which could have been achieved with one of the clustering accuracy method from [**?**].
- The samples all originated from the same bacteria's DNA. Had i used DNA samples from a diverse origin, i could more confidently have certified the precision of **MM** and **MM**$\frac{1}{2}$.
- An extension to the above point, i wish the samples in the precision tests were of both DNA and RNA, to see whether there was a difference between when k and H were most precise for RNA, and when they were most precise for DNA. If there were such a distinction, it could be used to determine what value k and H should have depending on the input.

The speed tests were those that showed most potential, but needed a few more tests before they could have been conclusive. There were a few things i would have done differently here too, such as:

- I had mentioned that the `Silva` samples and `Actino` samples had a different average length of sequences. While this was true, the single experiment was not sufficient to test the hypothesis. I would have liked to test the speed on more samples, and for each of these samples to assure that they had the same number of sequences, but a different average length of sequences. The samples should not differ too much in type, maybe even belong to the same bacteria. Thereby, i could have confirmed whether the average length of the sequences had a positive influnce on the speed of **MM** over `uClust`.
- I would also have liked to check whether there was correlation between whether the sample was DNA or RNA and the runtime of **MM** and `uClust`. The few experiments suggested that **MM** runs faster than `uClust` with DNA as input, but there was not enough data to cement this.

So, while the results showed some interesting findings, I would need more testing before i could properly prove that **MM** were a proper contender to `uClust`.