1 - Plan for subjects

These plans are made according to the 8 subjects that are mentioned in the Midway Report. They were written as i had to commence each subject, so that i had a plan of actions to tackle the subjects.

1. (1) Test of universal hashing schemes
   Date: 22 April 2015.
   Goal: To test both hashing schemes, and see which is quickest.
   Procedure: I shall use the two hash functions i defined in my universal hashing section.

   For testing, i shall use three different fasta files of different sizes. These are to be found in the directory given by Sune, wherein a lot of fasta files are located.

   Before testing, i must program both hashing function. This should no be too difficult as i have a working prototype running, so all that has to be done is copy the working implementation of the first hash function and change is so that it applies to the other hash function.

   Finally i shall have my java program write into a file the results of each run, so that i can create a table of the runtimes of each hashfunction.

   When the table is complete i can begin commenting on my results and see which hashfunction to use.

   Result: Done on April 24. Found that the multiply-shift was a tad quicker. Had to use randomized k-mer transformations, for the reason given just below. **Bugs**: Found a problem of memory when the number of sequences was at 5 mio. The k-mer transformations took too much space. This is to be expected in a single threaded implementation, and should be circumvented by using Map reduce.

2. (4) Test of k-mer sizes
   Date: 24 April 2015
   Goal: To find which k-mersize is closest to the gold standard.
   Procedure: I shall use 5 fasta files with approximately 10000 sequences in it to perform tests on. This is because my calculation of the gold standard will be using levenshtein distance which is very slow and therefore the test files cannot be larger than this.

   When the gold standards are found, i shall perform tests using my prototype. These will be by performing tests for 10 different k-mer sizes and see whether there is a pattern. If not, either the size of the file must be changed, or other k-mer sizes must be tested.

   When the data makes sense, i shall make a table of it, and comment on this table.