You are given a data set consisting of DNA sequences (the file is available [here](https://drive.google.com/open?id=1cIHHhgrM0jNGhULu2s-jjhAporHm39DY)) of the same length. Each DNA sequence is a string of characters from the alphabet ‘A’,’C’,’T’,’G’. The sequences are feature vectors and characters are features. The data set is stored as a fasta file, which is essentially a text file which stores the data in the following form:

>Sequence1

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

>Sequence2

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

>Sequence3

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

Here each line starting with ‘>’ symbol contains the name of a sequence followed by the sequence itself in the next line.

1) Propose how to convert feature vectors of characters into feature vectors of numbers. Justify your approach. Write a script which converts DNA sequences from our dataset into numerical format.

2) Build a visualization of the DNA data using two methods:

a) Multidimensional scaling with 2 dimensions. An input distance matrix should be the matrix of pairwise Hamming distances between sequences.

b) Principal component analysis. Reduce the numerical data matrix built in 1) to 2 dimensions and plot sequences using 2 obtained features as coordinates.

Which method do you think produced the better result? Explain your answer.

3) Cluster DNA sequences by applying k-means algorithm to 2-dimensional projections of the data obtained in 2). Identify the best value of k and justify your selection. Which projection method (PCA or MDS) result in better clustering?

You may use library functions to read data from the file and perform multidimensional scaling. For principal component analysis, you may use library functions to calculate eigenvalues and eigenvectors. K-means clustering should be implemented from scratch. Your submission should contain:

* The code of your scripts/programs
* Visualization plots for MDS and PCA with different clusters highlighted in different colors.
* Pdf or word file with your answers to the questions