Documentation for Homework 3

My scripts for this assignment consist of 2 python files:

* **data\_processing.py**
* **autoencoder.py**

First file can be run by the command:

*Python dataprocessing.py*

and will preprocess the 3 fasta files inside the folder fasta\_files. The output from running this file is numpy files containing the probability matrix for the different kmers for each sequence. It will generate 3 numpy files, 1 for each family of sequences.

The second file can be run by the command:

*Python autoencoder.py*

It will load the preprocessed data and merge them together in order to create the data used in the model. Then the file will create the autoencoder model, train the model on 75% of the data and then test it on the remaining 25%. Afterwards it prints the encoded vector in the prompter but since it is not possible to see the whole vector it also writes the result to a separate file: *“result.txt”* that will be created in the same folder.

A graph over the loss will be displayed in a separate window and we can see that the model has reached convergence.