Exercise 1 – Report

Antigen Discovery for SARS-CoV-2 (“Corona”) Virus Vaccine

# Data Handling

Each data item is a 9-lengthed sequence of characters, and all of the characters are in a vocabulary of 20 characters. We decided to represent each sequence to 9\*20 vector – concatenation of nine (one for character) one-hot encoding vectors, each one is of size 20. This seems like a natural choice because we do not have prior assumptions about the data, we want to keep the order and we do not know if there is any connection between some of the characters.