1. First, we will use the string kernel function for our kernel. Recall from class that given two strings s and t, the string kernel $K_p(s, t)$ is the number of substrings of length p that are common to both s and t, where a string that occurs a times in s and b times in t is counted ab times. For this problem, use p = 3, p = 4 and p = 5. Write down the training and test errors of kernel perceptron for p = 3, 4, 5 on this dataset.

	Training Error	Test Error
p = 3	0.0124	0.0409
p = 4	0.0069	0.0264
p = 5	0.0069	0.0343

2. Next, repeat Part (1) with a slight modification of the string kernel, $M_p(s, t)$. Given two strings s and t, the modified string kernel $M_p(s, t)$ is the number of substrings of length p that are common to both s and t, where a string that occurs a times in s and b times in t is counted only once. What are the training and test errors for this kernel for p = 3, 4, 5?

	Training Error	Test Error
p = 3	0.0127	0.0541
p = 4	0.0074	0.0290
p = 5	0.0069	0.0343

3. Find the two coordinates in w with the highest positive values. You should be able to do this without explicitly computing all the coordinates of w. What are the substrings corresponding to these coordinates? These coordinates correspond to those substrings whose presence most strongly indicates that the protein belongs in the family.

The substrings corresponding to the coordinates with the highest positive values are “WDTAG”, “DTAGQ”, “LFLNK”, “GKSSL”, and “KVGPD”. All of them have coordinate value of 3.