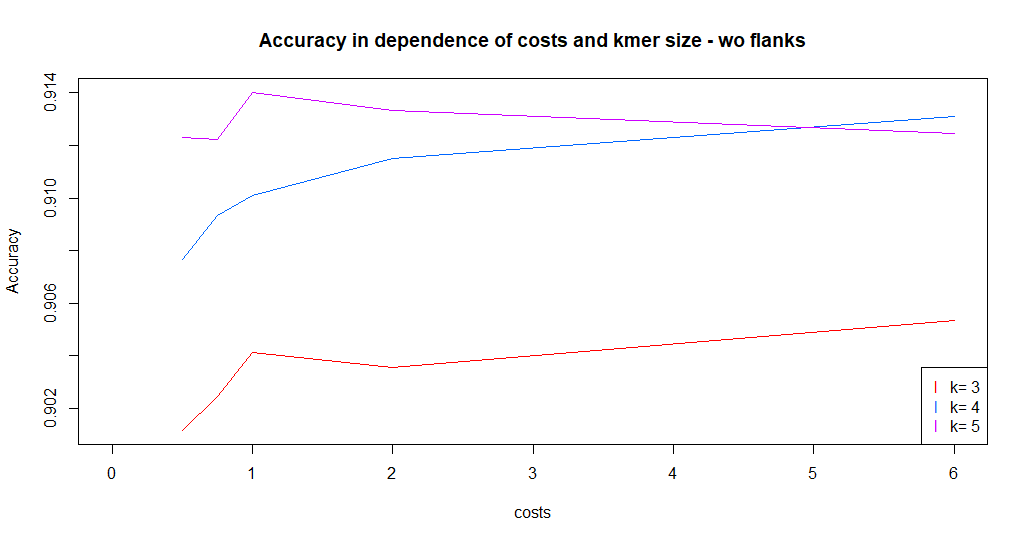
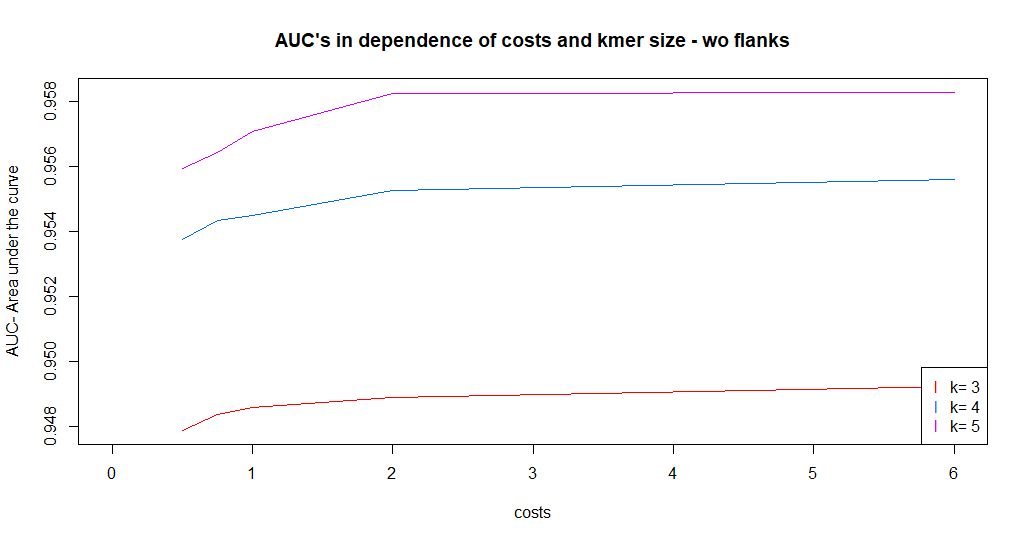
**Exercise 5  
Ben Wulf , Amnon Bleich, Lie Hong**

**Task 1**

For our assignment, we use the PUM2 dataset.

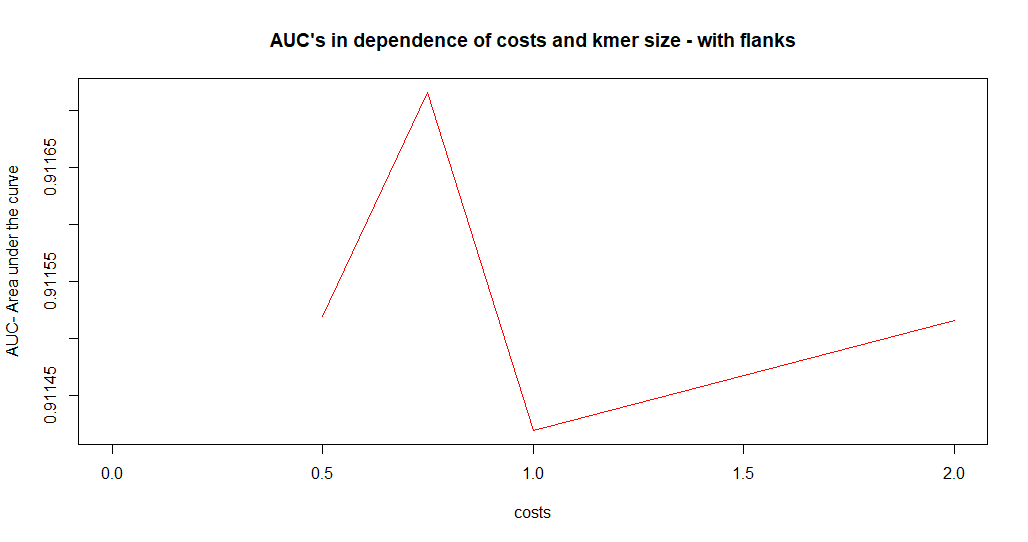
**Task 2/3**

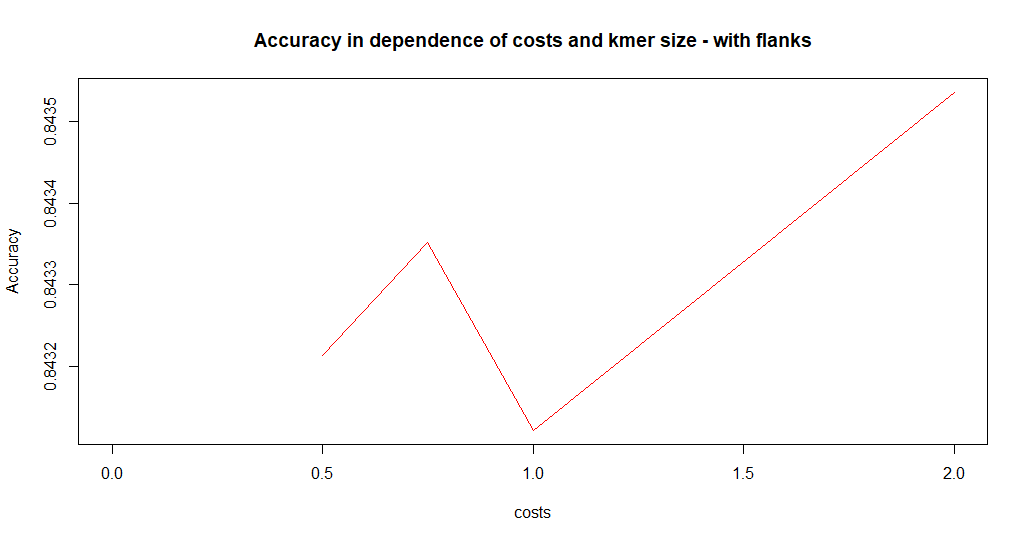
We perform the SVM training for different kmer (3,4,5) sizes and different c (0.5,0.75,1,2) for the binding sides. For the task we use the Kebabs package and for the validation we use  
the build in 10-fold cross validation. We use Kebabs, because we were not able to run   
Kernlab successfully. It always crash while blowing up the Ram.  
We reach the best accuracy with a kmer size of 5 and cost 1. The best AUC was reached by using a kmer size of 5 and a cost of 6. We think cost of 6 will lead to overfitting in a general case. So in a real case we would use a kmer size of 5  
and cost of one or two.



**Task 4**

For time reasons, we decide to use for the dataset with flanks only a kmer size of 3 and the cost steps 0.5,0.75,1,2





We can easily see that using flanks reduce the accuracy for 10% and the AUC for 4%. That means that it is not usefull to include the flanks.

**Task 5**

The top kmers with the biggest influence on the classification are: AGA GTT GTC

-4.05114273 -3.93739225 -3.88236865