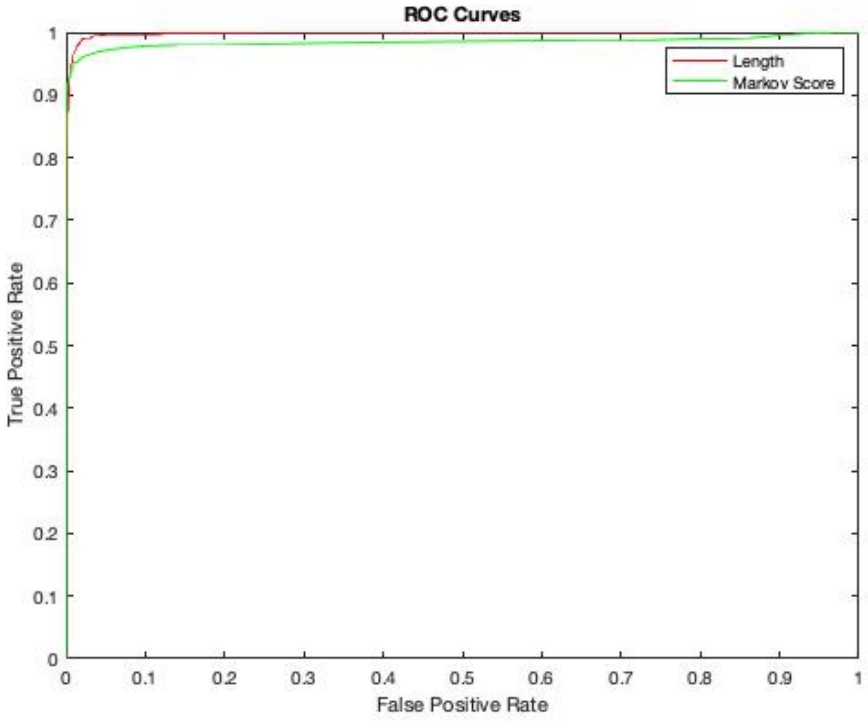
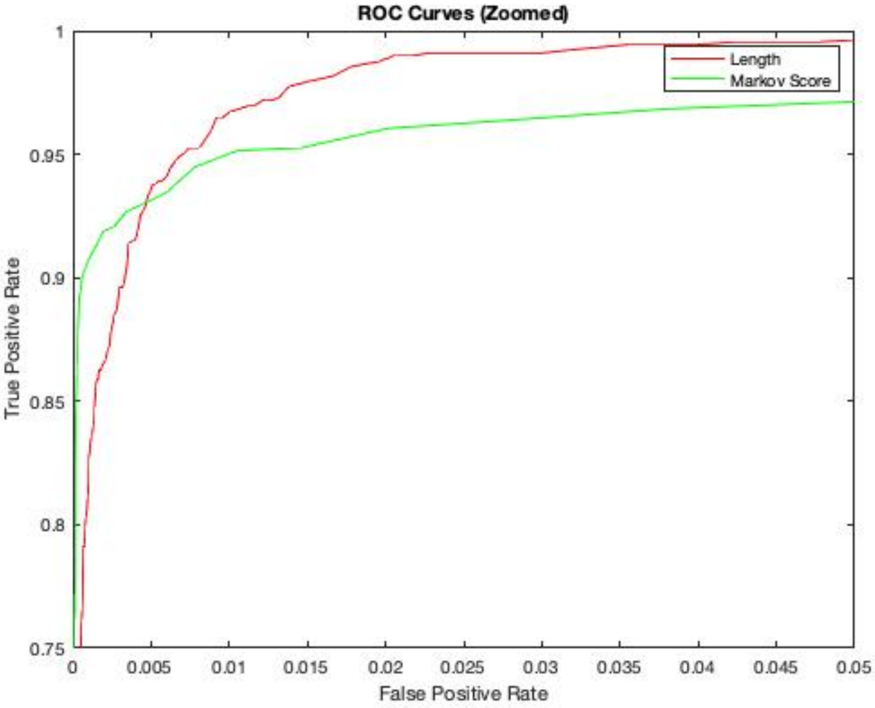
```
--------------------------------------------------------------------------
--------------------------     TASK 1A:     ------------------------------
--------------------------------------------------------------------------

READING FRAME 1:
Total Number of ORFs:  37897
Summary of First ORF:  start: 1,        stop: 36,       length: 36,
known match: false,   MM Score: 1.209843
Summary of Last ORF :  start: 1664953,  stop: 1664964,  length: 12,
known match: false,   MM Score: 0.387672


READING FRAME 2:
Total Number of ORFs:  38772
Summary of First ORF:  start: 2,        stop: 94,       length: 93,
known match: false,   MM Score: 2.160091
Summary of Last ORF :  start: 1664921,  stop: 1664968,  length: 48,
known match: false,   MM Score: -1.319054


READING FRAME 3:
Total Number of ORFs:  38530
Summary of First ORF:  start: 3,        stop: 5,        length: 3,
known match: false,   MM Score: 0.000000
Summary of Last ORF :  start: 1664964,  stop: 1664969,  length: 6,
known match: false,   MM Score: 1.490284


--------------------------------------------------------------------------
----------------------     TASK 1B, 1C & 1D:     -------------------------
--------------------------------------------------------------------------

SHORT ORFs (length < 50):  72771
LONG ORFs (length > 1400): 118
POS-STRAND CDSs in GENBANK:    892


--------------------------------------------------------------------------
--------------------------     TASK 1E:     ------------------------------
--------------------------------------------------------------------------

P(T|AAGxy):

          A          C          G          T
A     0.210130   0.183453   0.202359   0.353046
C     0.332268   0.096153       0.24   0.327823
G     0.266078   0.189189   0.131756   0.605555
T     0.407713   0.193877   0.247272   0.256144


Q(T|AAGxy):

          A          C          G          T
A     0.396475   0.268041   0.410256   0.357142
C     0.450777   0.132530        0.5   0.329639
G     0.280821   0.194174   0.419354   0.375796
T     0.486013   0.194029   0.496124   0.362318
```

```
-------------------------------------------------------------------------
--------------------------     TASK 1F:     -----------------------------
-------------------------------------------------------------------------

FIRST 5 SHORT ORFs SUMMARIES:
start: 1,        stop: 36,       length: 36,      known match: false,   MM
Score: 1.209843
start: 9,        stop: 20,       length: 12,      known match: false,   MM
Score: -0.838218
start: 24,       stop: 32,       length: 9,       known match: false,   MM
Score: 1.064601
start: 40,       stop: 51,       length: 12,      known match: false,   MM
Score: 2.084566
start: 55,       stop: 72,       length: 18,      known match: false,   MM
Score: -1.963672


FIRST 5 LONG ORFs SUMMARIES:
start: 17619,    stop: 19229,    length: 1611,    known match: true,    MM
Score: 166.008491
start: 33626,    stop: 35245,    length: 1620,    known match: true,    MM
Score: 207.933975
start: 42725,    stop: 45109,    length: 2385,    known match: true,    MM
Score: 258.426726
start: 74592,    stop: 76010,    length: 1419,    known match: true,    MM
Score: 137.311860
start: 76820,    stop: 78481,    length: 1662,    known match: true,    MM
Score: 202.510528
```

```
----------------------------------------------------------------
-----------------------     TASK 2:     ------------------------
----------------------------------------------------------------
```

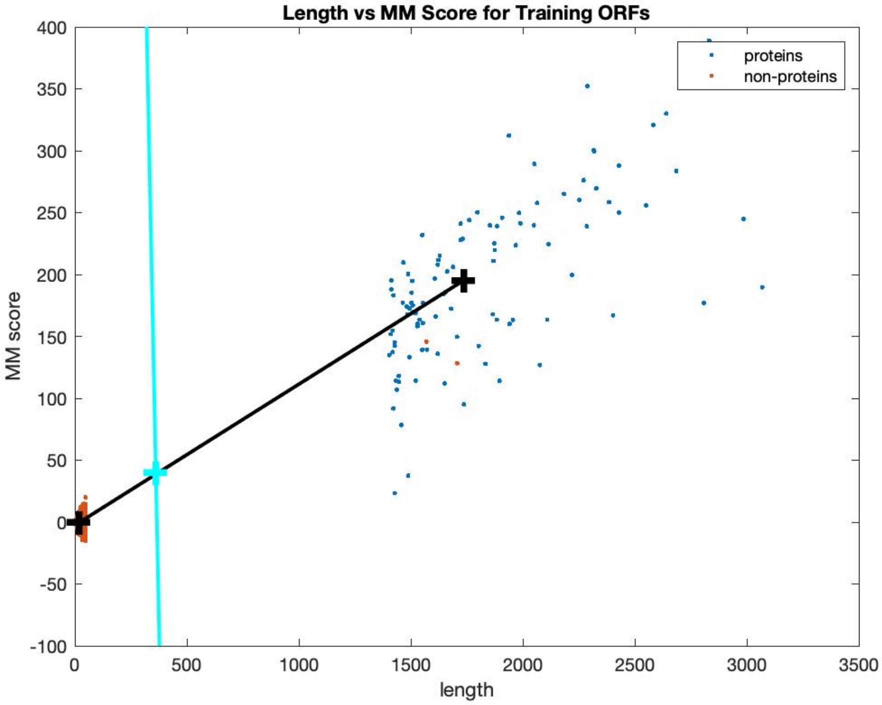

ROC Curves
Length
Markov Score
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


ROC Curves (Zoomed)
Length
Markov Score
True Positive Rate
0.75
0.8
0.85
0.9
0.95
1
False Positive Rate
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0.05

```
-------------------------------------------------------------------------
---------------------------     TASK 3:     -----------------------------
-------------------------------------------------------------------------

LENGTH THRESHOLD:   Length=414,           TPR=0.8024,       FPR=8.1355E-4

-------------------------------------------------------------------------
---------------------------     TASK 4:     -----------------------------
-------------------------------------------------------------------------

MM SCORE THRESHOLD: MMscore=34.3448,      TPR=0.8058,       FPR=1.3996E-4
```

```
-------------------------------------------------------------------------
---------------------------     TASK 5:     ----------------------------
-------------------------------------------------------------------------
```

Training Data:



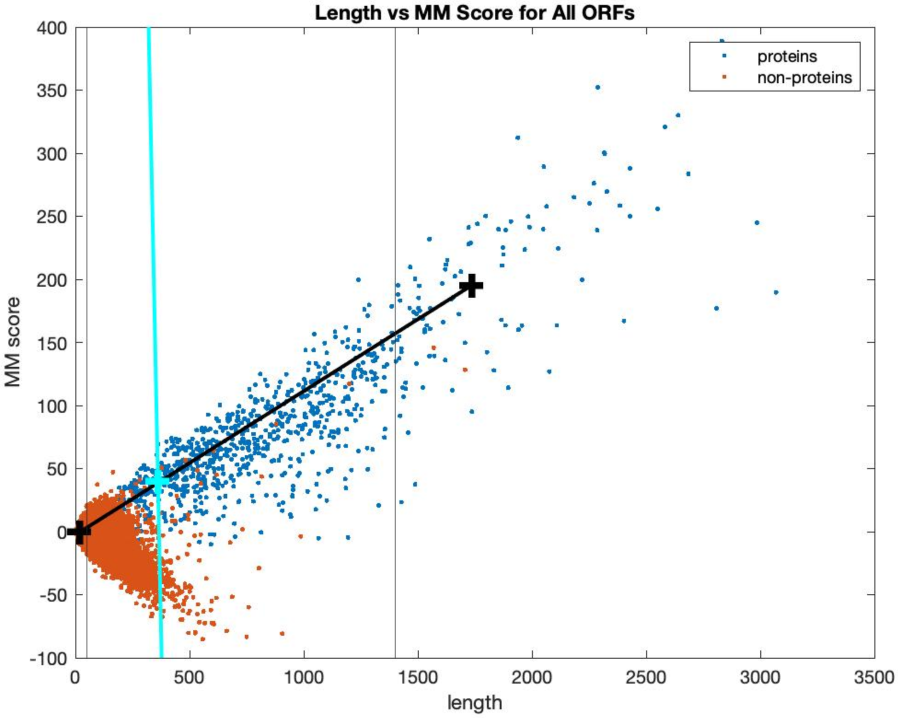
The two + signs are the medians of the short ORFs and long ORFs:

```
short data // median length  : 18.0
short data // median score   : -0.503917454067409
long data  // median length  : 1737.0
long data  // median score   : 195.32885752162156
```

The blue line represents a separation perpendicular (to the black line connecting the short data and long data) positioned 20% between the short median and long median.

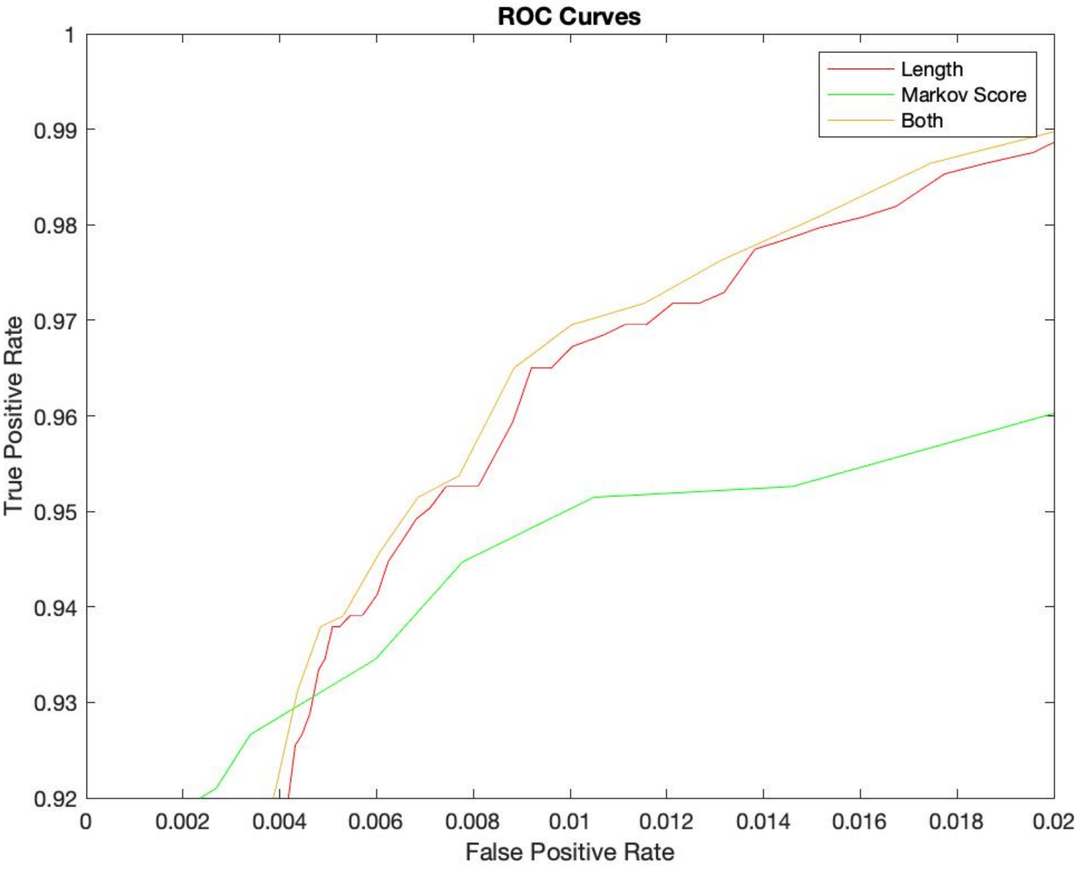
Slope of black line:
Slope of blue line:

```
Testing Data:
```



```
For the above data (blue line positioned at 20%), we had the following
performance:

TPR = 0.8465011286681715
FPR = 0.0012159596896241021
```

ROC CURVE:
The yellow line represents the classifier graphed above. It was made by shifting the blue line between the medians. For comparison, the length and Markov score ROC curves have been included.



Using the ROC data, I calculated that the 80% TPR threshold is as follows:

```
        percentage=0.23, TPR=0.801354401805869, FPR=7.523203835084374E-4
```

Reflection: From this project I was most surprised to see how effective length was as an ROC classifier. While the MM Score predictions were marginally better – its wasn't by much which was a fun reminder that computers and data can only do so much in terms of computations. I really enjoyed working on this and being able to do the graphs (especially for the linear classifier) helped me understand the basics of ML way better than most other examples I have encountered. I really appreciated this class as a whole – thanks for a great quarter!