Assignment 5

Krishan Subudhi : ksubudhi@uw.edu

Student Number : 2040900

Date : 12/06/2020

1

- a) for each of the three reading frames, the number of ORFs you find, and the summary of the first and last of each Markov model parameters: k = 5, pseudo_count=1

  Reading Frame : 1 Number of ORFs = 35200 Summary of the first and last :

	index	start	end	length	frame	isCDS	scores
0	0	1	36	36	1	False	1.181405
35199	35199	1664968	1664970	3	1	False	NaN

```
Reading Frame :  2
Number of ORFs =  35933
Summary of the first and last :
```

	index	start	end	length	frame	isCDS	scores
0	35200	2	94	93	2	False	2.557823
35932	71132	1664921	1664968	48	2	False	-1.381238

```
Reading Frame :  3
Number of ORFs =  35686
Summary of the first and last :
```

	index	start	end	length	frame	isCDS	scores
0	71133	3	5	3	3	False	NaN
35685	106818	1664964	1664969	6	3	False	1.869675

- b) The total number of short ORFs (length less than 50)= 81738
- c) The total number of long ORFs (length greater than 1400) = 118
- d) The total number of simple plus strand CDSs found in GenBank = 848

- e) P(T | AAGxy) and Q(T | AAGxy) for the 16 possible combinations of x,y in A,C,G,T :

  Foreground T|AAGxy counts for P(T | AAGxy) =

	A	C	G	T
A	307	51	223	394
C	104	15	12	119
G	211	42	39	218
T	148	19	68	198

  Background T|AAGxy counts for Q(T | AAGxy) =

	A	C	G	T
A	90	26	48	95
C	87	22	15	119
G	41	20	26	59
T	139	39	64	175

- f)Summary data for the first 5 short ORFs

	start	end	length	frame	isCDS	scores
0	1	36	36	1	False	1.181405
71133	3	5	3	3	False	NaN
71134	9	20	12	3	False	-0.463650
71135	24	32	9	3	False	0.890789
1	40	51	12	1	False	2.758633

- f) Summary data for the the first 5 long ORFs

	start	end	length	frame	isCDS	scores
71526	17619	19229	1611	3	True	166.509569
36031	33626	35245	1620	2	True	208.460996
36169	42725	45109	2385	2	True	258.774328
72661	74592	76010	1419	3	True	138.186040
36888	76820	78481	1662	2	True	202.815509

Extra:

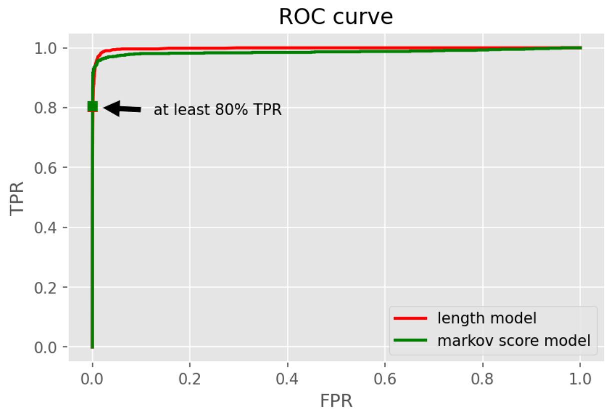
CDSs without stop codon at the end

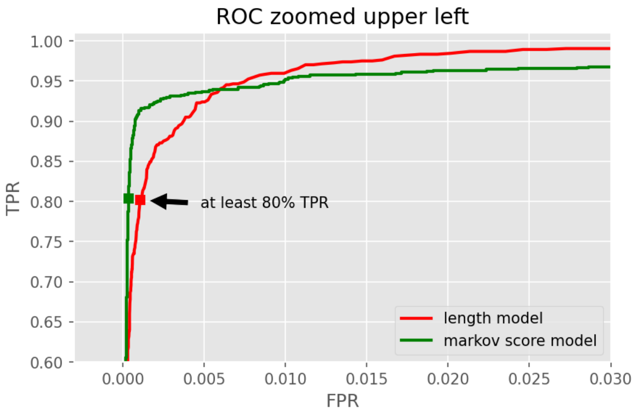
Some CDSs found in gff files did nto have stop codons at the end

```
ORFS which are CDS = 842, Total CDSs = 848
755683 TCG TCGTTA
742666 CCA CCACTG
754669 AAA AAATCA
1563085 TTT TTTTGG
753619 GAT GATAAA
15774 GGT GGTTCG
```

2. ROC

1. Generate a single plot showing ROC curves with respect to
    1. length threshold, say in red, and
    2. Markov model score, say in green, using the full 0-1 range for both axes.
2. Additionally, (c) generate such a plot "zoomed-in" to the upper-left corner to show the crossover between the two curves.
3. Also calculate and show Area Under the Curve (AUC) for each curve.





AUC

Length threshold AUC = 0.9978549668337447

Markov score AUC = 0.986286204546829

3

If your only option was to predict based on an ORF length threshold, what is the ~minimum~ maximum threshold that would achieve a true positive rate of at least 80%, how many true positives and how many false positives would you see using that threshold? Optionally, plot this point on the ROC curve above.

Prediction\Truth	Positive	Negative

Prediction\Truth	Positive	Negative
Positive	True Poistive	False Positive
Negative	False Negative	True Negative

```
TPR = TP/P = TP/(TP+FN)
FPR = FP/N = FP/(FP+TN)
```

Maximum length threshold for at least 80% TPR = 432

using that threshold true positives = 675 and false positives = 102

Point shown on the ROC curve.

4

If your only option was to predict based on a log Markov model score threshold, what is the minimum maximum threshold that would achieve a true positive rate of at least 80%, how many true positives and how many false positives would you see using that threshold? Optionally, plot this point on the ROC curve above.

Maximum length threshold for at least 80% TPR = 35.6518

using that threshold true positives = 676 and false positives = 34

Point shown on the ROC curve.

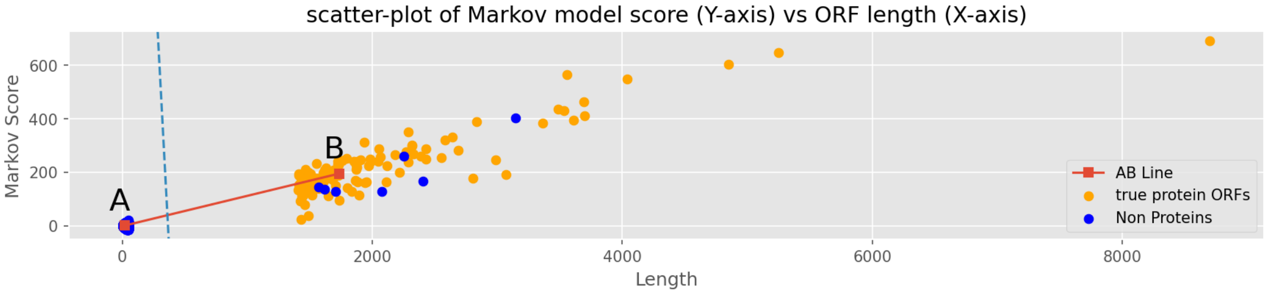
5. Flashbulb Classifier

1. Generate a scatter-plot of Markov model score (Y-axis) vs ORF length (X-axis) for each long and each short ORF.

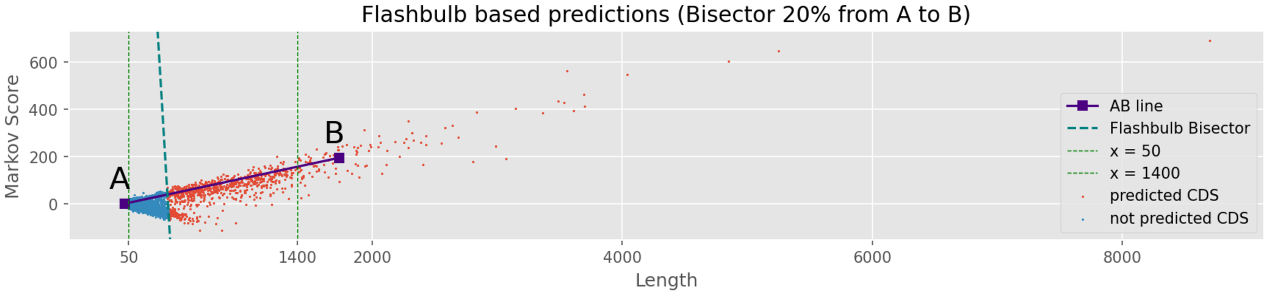
2. Color points according to their status wrt "simple plus strand CDSs" from GenBank (true protein ORFs: orange; non-proteins: blue)

3. Summarize the short ORFs by the single point that falls at the median x, median y of the ORFs of length < 50. Call this point A. Likewise, summarize the long ORFs by the single point that falls at the median x, median y of the ORFs of length > 1400; call this B. Overlay your plot with some visually distinct symbol at A and B, and connect them by a straight line segment.

4. Q: Also draw a straight line perpendicular to this line segment and crossing it at x = Ax + 0.20 * (Bx - Ax), i.e., 20% of the way from A to B. Calculate the equation of this line.



5. Q: Make another scatter plot, like the one requested at the start of this step, including the A-B line segment and perpendicular line at 20% (as previously calculated, i.e., just based on the training set), but this time plot points for all ORFs, not just the training ORFs. Add thin vertical lines at x=50 and x=1400



6. Q: How well does this work? Find its associated True Positive and False Positive counts and rates (on the set of all ORFs, not just the short/long training set).

   A: Using points above/to ther right the threshold line as +ve and below/to the left as negative,

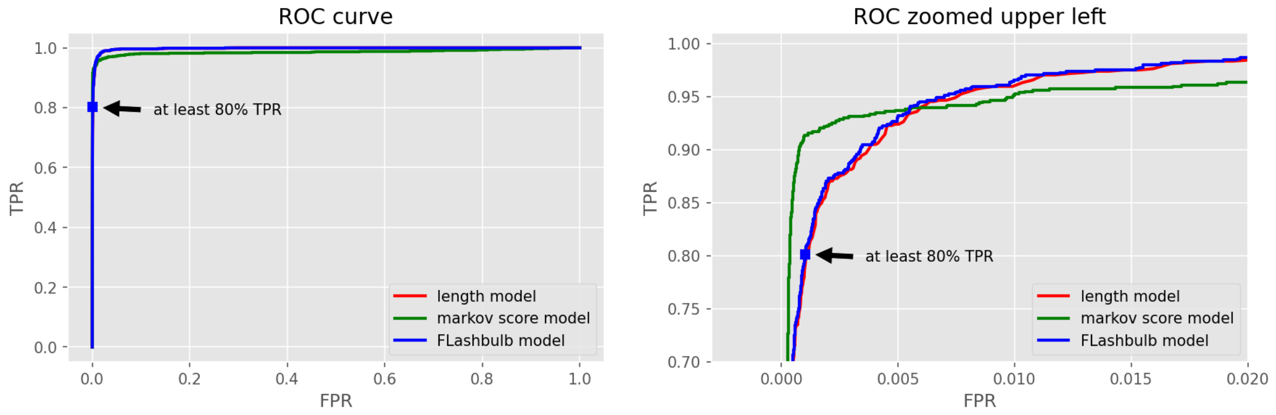
   true positives = 723 and false positives = 168, TPR = 0.8586698337292161, FPR = 0.0017317802288423874

7. Q: Varying that "20%" threshold from minus infinity to plus infinity, i.e., sliding a line parallel to the original 20% line across the plane, will give different tradeoffs between false positives and false negatives. Add the corresponding ROC curve to the graph from step 2 (use a different color), calculate its AUC, and for an 80% true positive rate, calculate the number of false positives (as in steps 3-4) (and optionally plot this point).

   A: I used the formula `y-mx-c` output as the flashbulb model score, then ROC curve is plotted using different thresholds for the score. For points `x,y` These scores are actully difference between the `y intercept` of a line with slope `m` passing through `x, y` vs the `y intercept` of line passing through 20% threshold point.

   Since slope is constant while calculating the intercept difference , This will be equivalent to sliding the flashbulb line explained in the assignment question.

   This method also avoids calculating different equations for the sliding bisector and then measuring distance of every point from those interceptors.



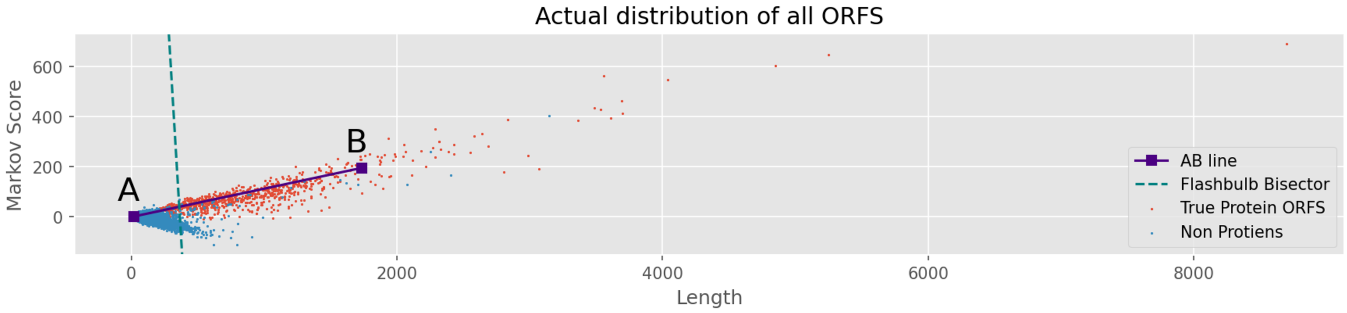
AUC

Flashbulb model AUC = 0.9979613728388557

For 80% TPR, true positives = 675 and false positives = 99

How to improve the results?

The actual true CDS vs false CSD distribution is shown in the following graph. The flashbulb line which sets the threshold clearly is not the best classifier. More precisely, the slope does not seem to be accurate.



So here are few suggestions for improvement.

1. Use a different slope. Perceptrons or gradient descent or even simple grid search on slope can be used to come up with the best slope.
2. Use a non linear boundry. Again, neural networks can be used to create a better boundry.
3. Use annotated genes to build the markov model instead of relying on long ORFS. That way the markov scores will be more accurate.
4. Increase training data.
5. Try with higher/lower order markov models. Tuning k can help the results. As a thumb rule, try higher k if more data available else experiment with lower k.
6. Tune pseudo count. Find number of kmers having zero probability without pseudo count. If the number is high, reduce pseudo count or reduce k.
7. Remove stop codons from -ve training data.