CSEP 527 HW 2- MEME Report

Isabelle Lee, (ID: 1423762)

November 13, 2020

1 How to run

Below command generates an output/output.txt file containing results and corresponding plots.

```
python src/meme.py -tr data/hw3-train.fasta -tt data/hw3-test.fasta
```

To download logomaker, please run

```
pip install logomaker
```

2 Entropies for iterations 0 to 10

For full results, please take a look at output/output.txt file. Looking at the table, we can see that the entropy increases as the iterations increase. This is due to information gain for each motifs increasing as we iterate. At the end of 3 iterations, 3 motifs were picked; A is motif 10, B is motif 15, and C is motif 20. Finally, after 10 iterations, motif 10 was chosen as D.

3 Sequence Logos

For extra credit, I printed sequence logos using logomaker package. Below plot shows sequence logos using WMMs. Below is useful visualizing which motif would be highlighted by corresponding WMMs.

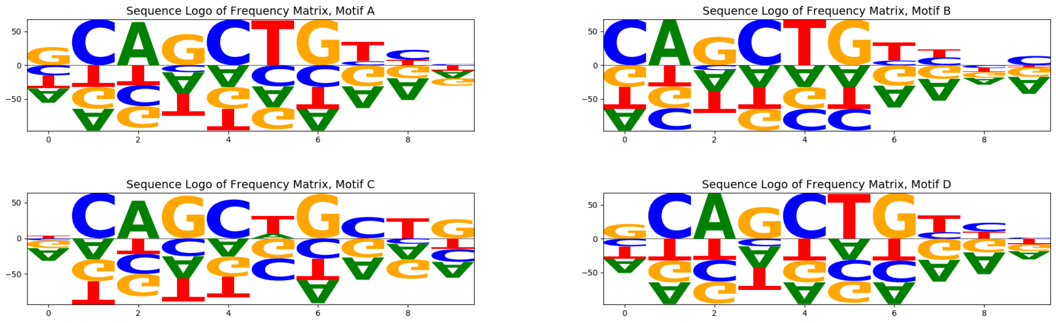


Figure 1: [Extra Credit] Sequence logos of Motifs A, B, C, D

4 Histograms for Starting Positions

Then, using the resulting WMMs, the histograms of starting positions were counted as shown below. For the most part, the chosen motifs picked position 51 or 52 as the starting positions.

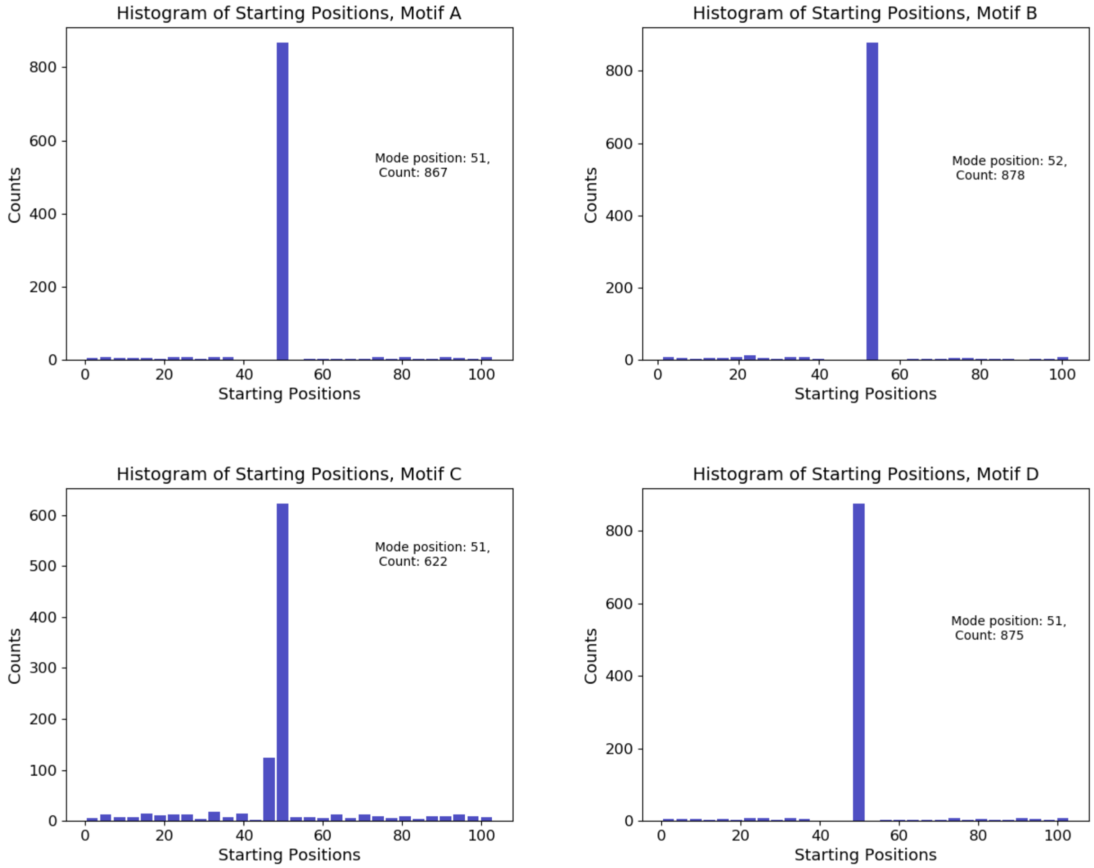
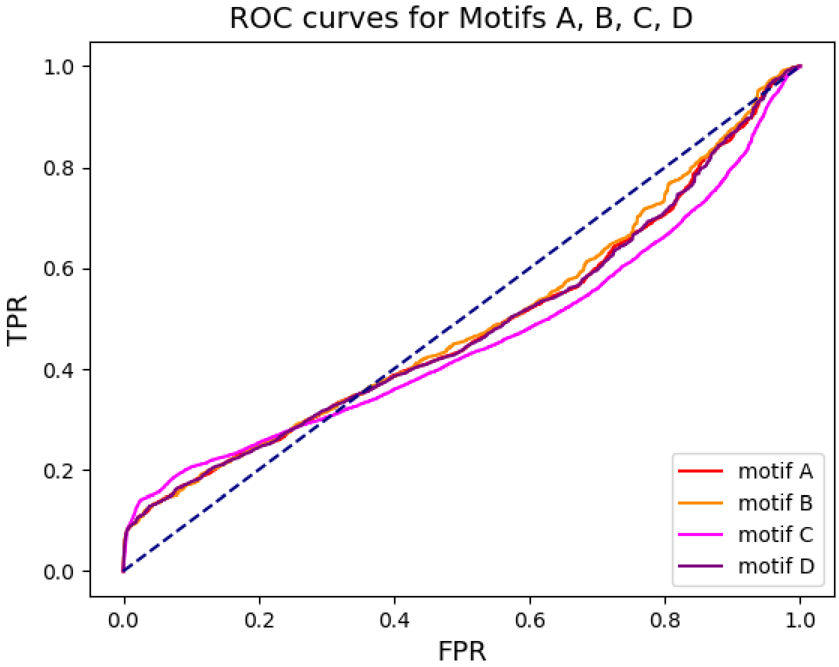


Figure 2: Starting position histograms for Motifs A, B, C, D

5 ROC curve

The scores from 4 results are used to plot ROC curves as shown below. It seems that lower cutoffs are optimal for maximizing true positive rates and minimizing false positive rates. For C, the best threshold was chosen at -44.684, with TPR of .361 and FPR of .4. The best threshold calculation results are also printed out in AUC scores section of output/output.txt.

I think there has to be a bug, since this ROC plot shows that the motifs did a poorer job at guessing the start position than random guesses..:(I suspect that there's an error in how I created the ROC plot or how I updated frequencies. I'm running out of time, so I'll turn in what I have.


ROC curves for Motifs A, B, C, D
TPR
FPR
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
motif A
motif B
motif C
motif D