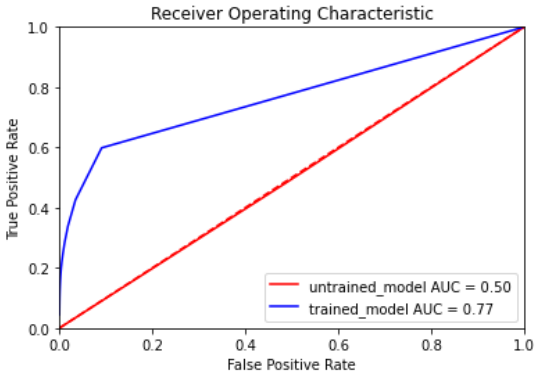
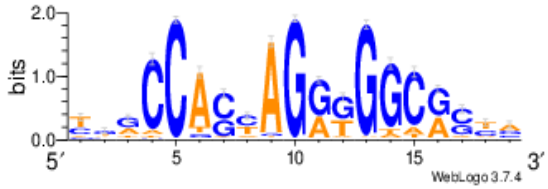
6135 - Assignment 1 - Practical Report

Ans 5.1



AUC-ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability.
The higher the AUC, the better the model is. The plotted ROC curve against TPR and FPR has AUC scores of 0.50 and 0.77 for the untrained and trained models respectively.
We can infer that the untrained model has almost no separability and is almost randomly predicting each class with half probability. Whereas, the trained model is indeed learning as the ROC curve for the trained model is nearer to 1 on the top left corner.

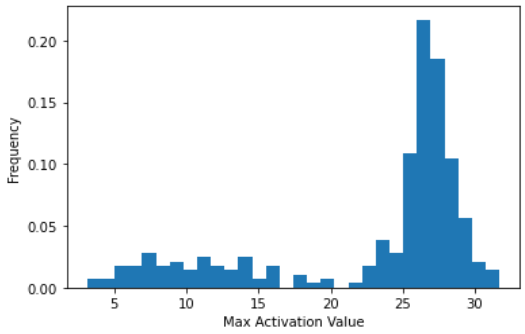
Ans 5.2



The normalized PWM (position weight matrices) for the given "jaspar" file of the CTCF motif is as follows. Also, the alignment of this motif generated by Biopython library's weblogo API is presented for better visualization.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A:	0.10	0.18	0.31	0.06	0.01	0.81	0.04	0.12	0.93	0.01	0.37	0.06	0.01	0.06	0.11	0.41	0.09	0.13	0.44
C:	0.32	0.16	0.05	0.88	0.99	0.01	0.58	0.47	0.01	0.00	0.00	0.01	0.00	0.01	0.81	0.01	0.53	0.35	0.20
G:	0.08	0.45	0.49	0.02	0.00	0.07	0.37	0.05	0.04	0.99	0.62	0.55	0.98	0.85	0.01	0.56	0.34	0.08	0.29
T:	0.50	0.20	0.15	0.04	0.00	0.10	0.01	0.36	0.02	0.00	0.01	0.37	0.01	0.08	0.07	0.02	0.04	0.44	0.06

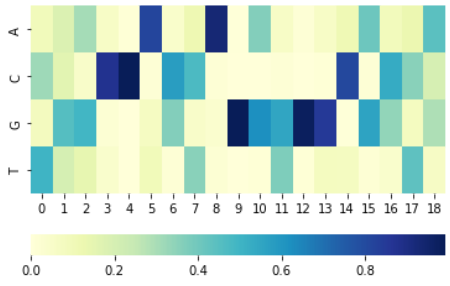
Ans 5.3



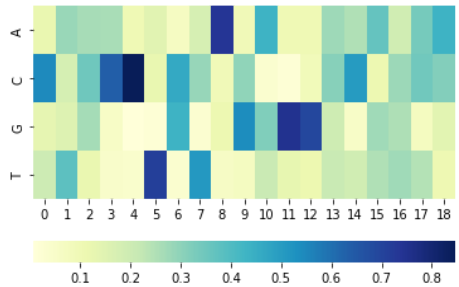
To convert each of the 300 filters of the first convolution layer, the first step is to get the maximum activated value across the entire test dataset.
These values are in the range 3.12 - 31.69. The histograms shown here are the distribution of all these 300 max activation values.

Ans 5.4, 5.5

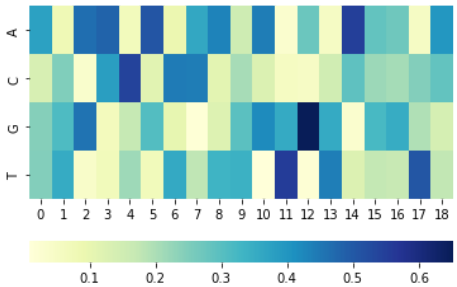
Using the unfold method of PyTorch the base-pair sequence that activates the filters more than 0.5 value of max is calculated. To check the similarity between CTCF and generated 300 normalized PWM from question 5.3, *Pearson Correlation Coefficient* is used. The most correlated PWM with CTCF has a value of 0.718 and the second most match PWM has a value of 0.6003. The following are their heatmaps which resembles quite similar to CTCF motif.



CTCF motif



Highest matched PWM



Second highest matched PWM