Assignment 3

Due Thursday, February 15th before noon (California time)

In this assignment, you will develop a Neural Network (NN) model using the gcPBM TF-DNA binding data you previously analyzed in Assignment 1. Your objective is to predict the TF-DNA binding affinity for the transcription factors Max, Mad, and Myc. Subsequently, we will explore another set of TF-DNA binding data created to study the impact of methylation on TF-DNA binding sites. For further details on the experiments and their findings, refer to the paper: (Kribelbauer et al., Cell Reports, 2017). In this part of the assignment, you are required to adjust your encoding method to incorporate 5mC (as 'M'), enabling the NN to predict TF-binding affinity in the presence of methylated DNA. This task will involve working directly with unaligned data.

1. You are tasked with developing a Neural Network (NN) model to predict the gcPBM TF-DNA binding affinity for three TFs: Max, Mad, and Myc, using a given DNA sequence, similar to your approach in Assignment 1. In this task, we will exclusively utilize 1-mer features. Implement 10-fold cross-validation to determine the average r-squared value. [2pt]
   Hint: Your NN should include a minimum of two hidden layers equipped with an adequate number of nodes. For the final layer, integrate a Dense(1) unit followed by a sigmoid activation function. Opt for mean squared error (mse) as your loss function, utilize the Adam optimizer, and apply the 'R2Score' metric to ascertain the r-squared value.

2. Compare the performance of the neural network using 1-mer features against the linear regression models that utilize both 1-mer and 2-mer encodings for Max, Mad, and Myc. Discuss your observations on their performance. Specifically, analyze why the neural network model with 1-mer data yields satisfactory outcomes, offering an explanation for this observation. [1pt]

3. Create a function capable of encoding 1-mer and 2-mer sequences, including sequences with 5-methylcytosine, denoted as 'M'. [2pt]
   Hint: For 1-mer encoding, use the following representations: A as 10000, C as 01000, G as 00100, T as 00010, and M as 00001. For 2-mer encoding, start with AA represented as 100000000000000000000000, proceed through combinations like AC as 010000000000000000000000, and conclude with MM as 000000000000000000000001.

4. Load and encode the EpiSelex-seq data for the TFs Atf4 (Atf4.txt) and Cebpb (Cebpb.txt). Apply the neural network model you developed in Question 1 and linear regression models using 1-mer and 2-mer features to predict their binding affinity to both methylated and unmethylated DNA sequences. Note that the binding data is unaligned. [2pt]

5. Compare the results and elucidate why the most effective model outperforms the other two in the context of this type of experimental data. [1pt]