Assignment 6

Due Thursday, March 21st before noon (California time)

Transcription factor (TF) binding to DNA is governed by two main mechanisms: base readout and shape readout. Base readout enables TFs to identify their DNA binding sites based on the DNA's physicochemical patterns, while shape readout allows TFs to recognize binding sites through the 3D structure of the DNA. Before the development of tools such as DNAshape, predicting the 3D structural features of DNA in a high-throughput manner was challenging. In this assignment, you will utilize the DNAshape tool to explore the significance of DNA shape in TF-DNA binding and assess its impact on the predictive capabilities, analyzing gcPBM data for the transcription factors Max, Mad, and Myc.

High-throughput in vitro data analysis:

1. Install the *Deep DNAshape* package from https://github.com/JinsenLi/deepDNAshape to obtain DNA shape features for the input files Max.txt, Mad.txt, and Myc.txt. Generate a feature vector for "1-mer" sequence model, a feature vector for "2-mer" sequence model and a feature vector for "1-mer+shape" model for each of the datasets corresponding to Mad, Max and Myc. [2pt]

2. Build L2-regularized multiple linear regression (MLR) models for "1-mer", "2-mer" and "1-mer+shape" features with 10-fold cross validation. Calculate and report the average R^2 (coefficient of determination) for each of these three models across the datasets of Mad, Max and Myc. [2pt]

3. Generate two plots for a comparison of two different models: one comparing the "1mer" vs. "1mer+shape" and another comparing the "2mer" vs. "1mer+shape" model. Make the plot in a manner similar to that presented in Figure 1(B) of Zhou et al. PNAS 2015. Briefly discuss what you have learned from the results. [2pt]

High-throughput in vivo data analysis:

4. Repeat the process outlined in Q1 for ChIP-seq data (ctcf_bound.fasta and ctcf_unbound.fasta) for the CTCF TF from Mus musculus. Build logistic regression models for the "1-mer", "2-mer" and "1-mer+shape" features, Plot the ROC curves for each model and calculate the AUC score for each curve. Briefly discuss what you have learned from the results. [2pt]