QBIO 481 Optional Assignment

Brandon Ye

2a. SELEX-seq is an *in vitro* experiment that is used to identify specific DNA or RNA sequences that bind to macromolecules such as proteins. PBM is an *in vitro* experiment that characterizes DNA-macromolecule interactions utilizing a microarray containing DNA sequences, where incubation of the protein in the microarray and measurement of binding affinity provides information about binding affinities.

2b. ChIP-seq is an *in vivo* experiment that associates certain regions of DNA associated with specific proteins by utilizing a specific antibody to precipitate protein-DNA complexes, which are then purified and sequences to map the DNA regions associated with the target protein.

2c. Both SELEX-seq and PBM provide control over experimental conditions at the sacrifice of measuring otherwise biologically relevant interactions that could be measured *in vivo*, such as in ChIP-seq.

5b. The points for Mad, Max, and Myc each lie above the y=x line in the 1-mer+shape axis in the plot. Therefore, the R^2 value of the model is increased with the addition of shape, indicating that a greater proportion of variance in our output variable can be attributed to input features. Consideration for shape can therefore improve model performance.

7b. The ensemble plots generated by plotShape() for the DNA shape parameters MGW, and ProT, Roll, and HelT reveal differences between the bound and unbound sequences, primarily in the parameters ProT and HelT. In ProT, we observe an absolute maximum near the center for the bound sequence, but only a relative maximum for the unbound sequence. In HelT, fluctuations in the mean value are more pronounced in the bound sequence compared to the unbound sequence.

8b. We observe that the AUC score for the 1-mer model curve is 0.858 while the AUC score for the 1-mer+shape model curve is 0.875. The AUC score for the 1-mer+shape model is higher. As we observed previously, consideration for shape improves the performance of the model.