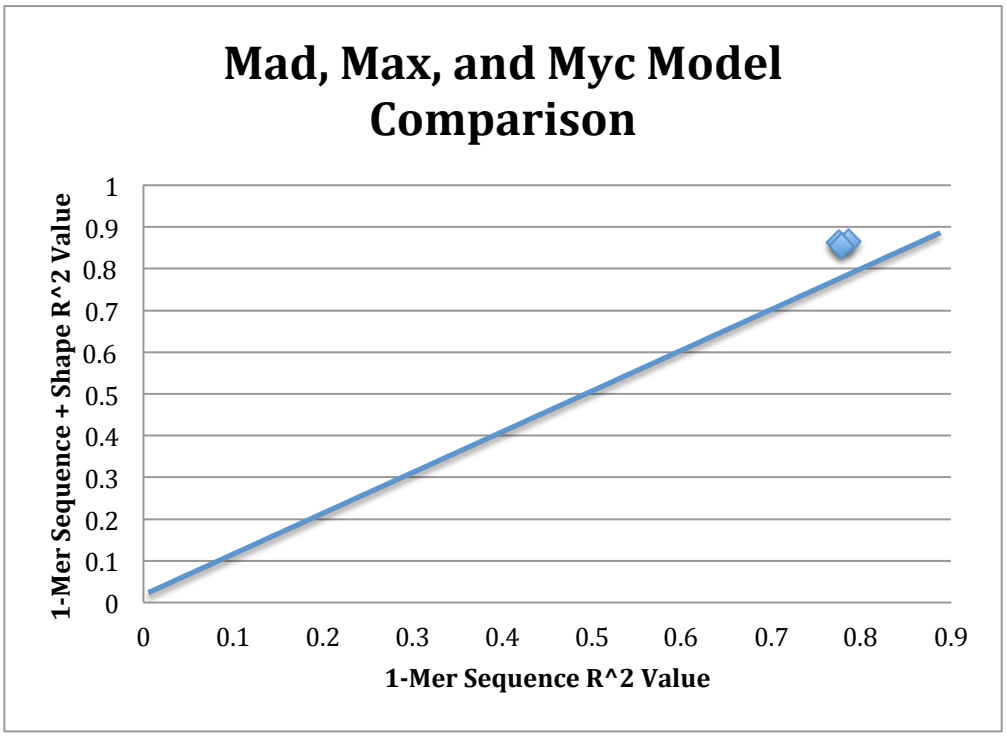
Anjali Chakravarti
BISC 481
Homework #3

1. Project created online.
2. A) High-throughput binding assays are used to determine which sequences correspond to the binding sites of a particular protein. In-vitro methods are done in the lab, outside of the cell. SELEX-Seq is the first of two in-vitro high-throughput binding assays. In this process, the proteins of interest are fixed to a plate. Then, multiple oligonucleotide sequences are generated and washed over the plate. Some bind to the protein if there is a sequence match, and the rest of the nucleotides are washed away. The sequences that bind are selected, amplified, and washed over the plate again in additional purification steps. Then, the selected sequences are purified and sequenced and the sequence of interest is determined.
   A protein binding microarray, or PBM, is the other in-vitro high-throughput binding assay. In this type of assay, the nucleotide sequences are placed in an array. The proteins are fluorescently labeled and washed over the array. The fluorescent tag allows for visualization of the sequences that the proteins bind to. Then, the sequences that are bound are visualized (fluorescence) and the nucleotides are sequenced.
   B) In-vivo methods are methods that involve the living organism. The in-vivo high-throughput binding assay is called ChIP-Seq or ChIP-chip. In both methods, the DNA in vivo is cross-linked with proteins. Then, the cells are lysed and the DNA is sonicated to break it up. The proteins are precipitated out of the solution (making use of the cross-linker proteins). Then, there is an optional immunoprecipitation step to separate out specific sequences. Finally, the DNA is un-crosslinked and labeled, and are either placed in a microarray (for ChIP-chip)
   C) The advantage of using in vivo methods is that the gene pool is exactly as
3. Downloads and installations.
4. The table of outputs is below:

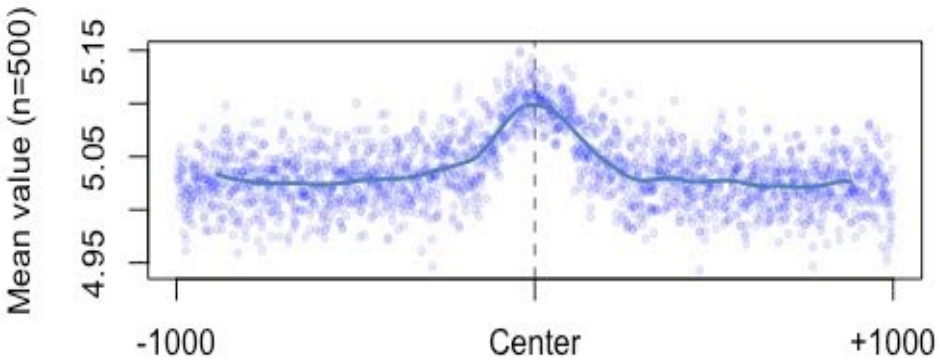
	R^2 value of 1-mer sequence model	R^2 value of 1-mer + shape sequence model
Mad	0.7752	0.8627
Max	0.7854	0.8643
Myc	0.7780	0.8546

5.

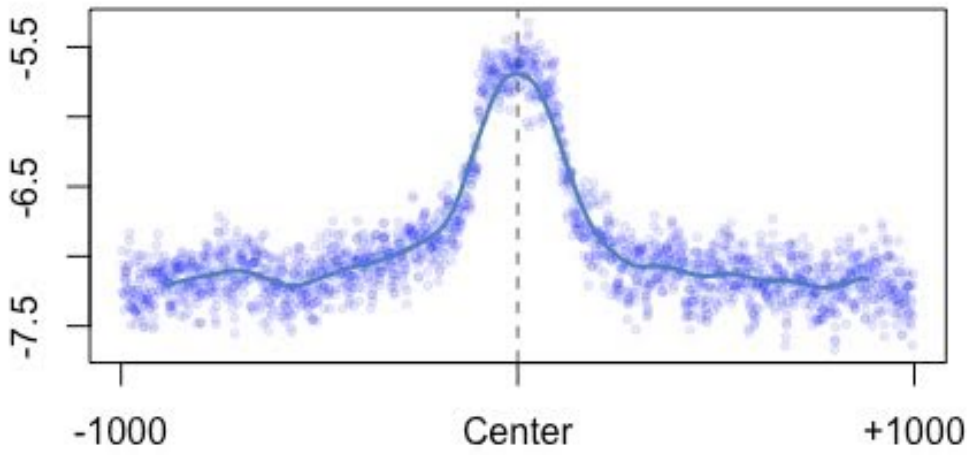


The binding affinity prediction of the model with the shape included is more accurate than the binding affinity prediction of models without it. This is clear based on the comparison of the R^2 values for each model, as done on this chart. Because all three data points are above the center line, it indicates higher accuracy for the shape included model.

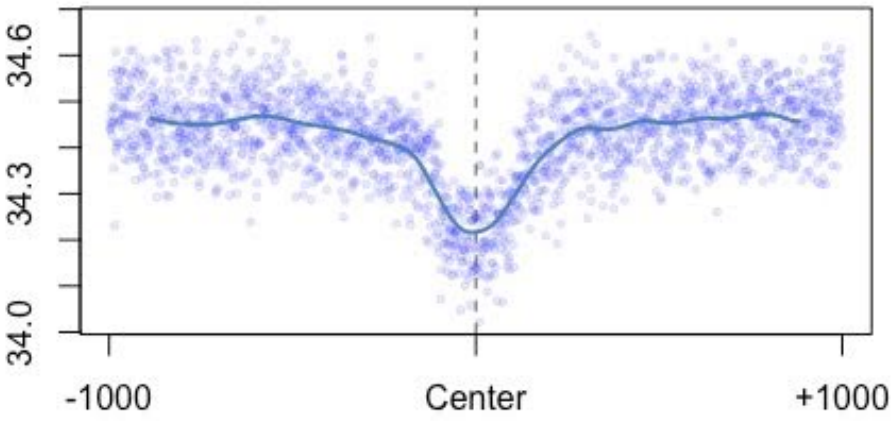
6. Downloads and Installations.
7. A) The Minor Groove Shape plot (mean val. n=500):



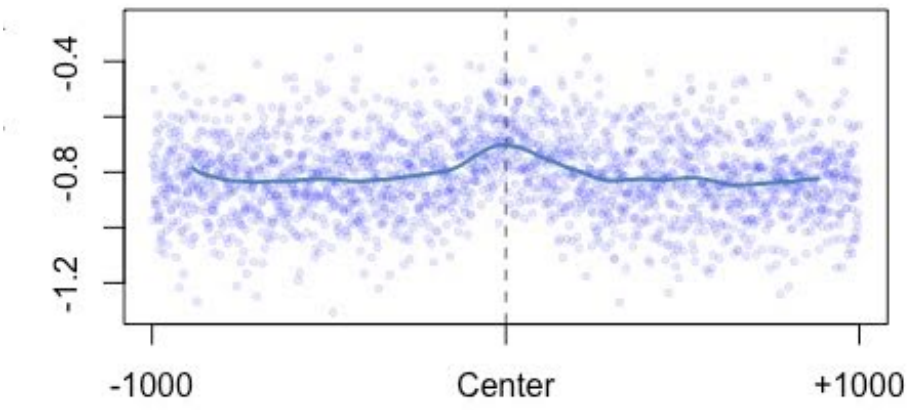
The Propeller Twist Shape Plot:



The Helix Twist Shape Plot:

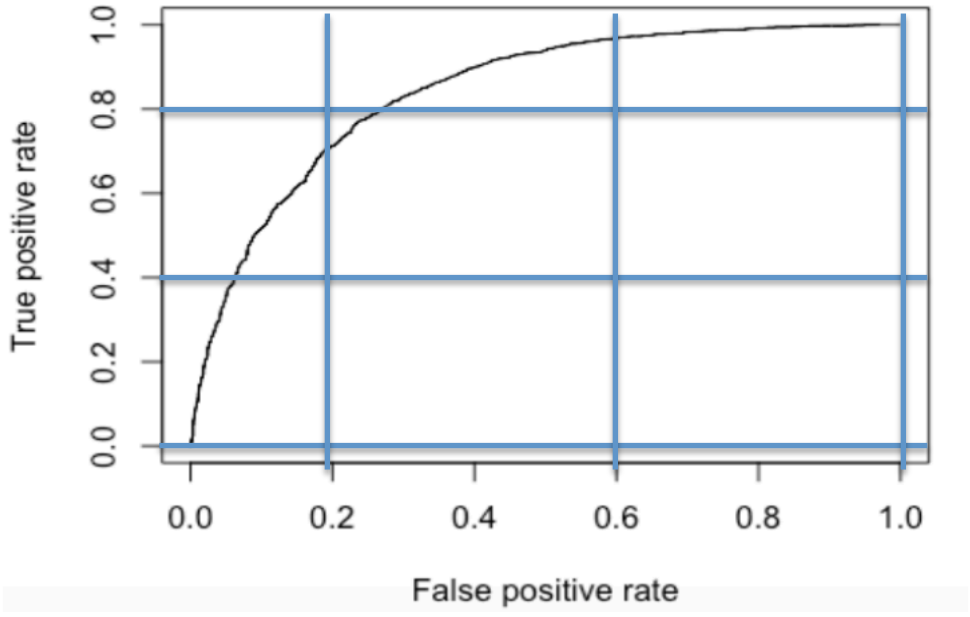


The Roll Shape Plot:

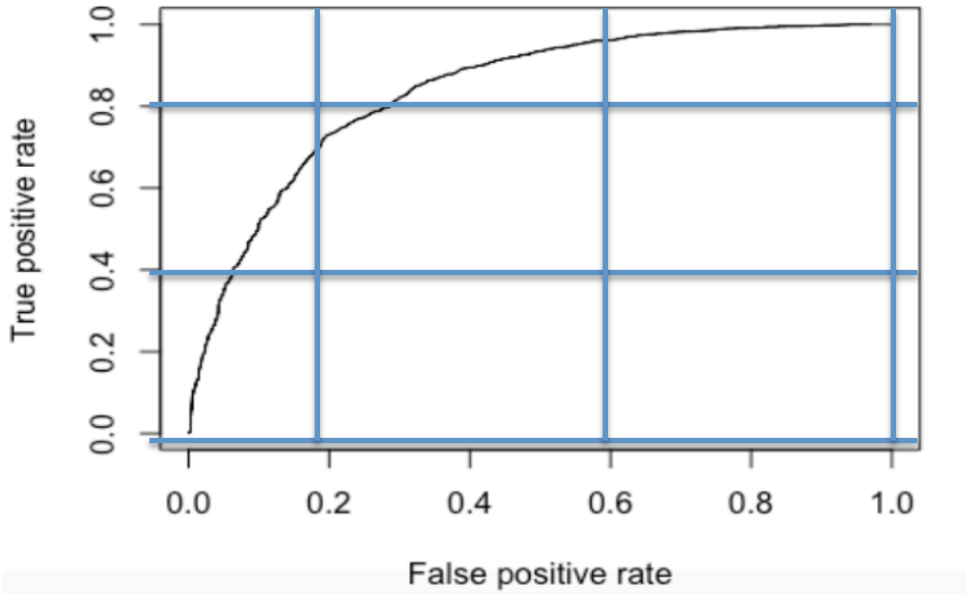


B) In the Roll, Propeller Twist, and Minor Groove sequences, the peak at the center. This means that the center of the pair is the widest part. This may be why those particular shapes are recognized, because that shape means there is more electrostatic focusing as the shape narrows toward the outside. For the helix shape, the valley at the center means that the shape is narrowest at the center, which means that there is less electrostatic focusing at that site and it is less conducive to shape readout.

8. A) The ROC Curve for 1-mer (AUC Score = 0.841):



The ROC Curve for 1-mer + shape (AUC Score = 0.837):



B) There is almost no difference between the AUC scores of the 1-mer model and the 1-mer + shape model (difference of 0.005). This means that the shape parameter has no effect on the results, so this protein uses sequence readout only to determine the binding region.