Brendon Cooper
5571607591

Remo Rohs' Assignment 2

1) https://github.com/nodnerb93/BISC577

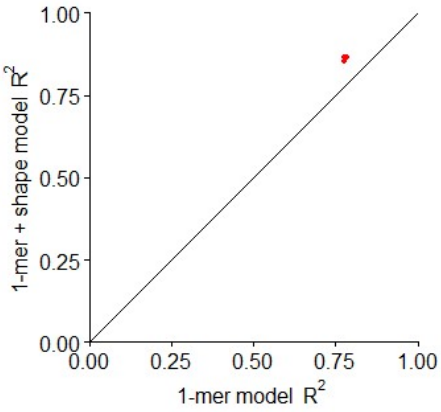
2) Hight-throughput binding assays
   a) SELEX-seq: Create large pool of random oligos and run over protein of interest bound to beads. Isolate oligos which bind to your protein of interest, amplify with PCR, then sequence all fragments with next generation sequencing. Relative affinity can be determined by comparing counts of each read. The higher the count for a specific sequence, the better its affinity.
   PBM: Affix predetermined oligos to chip then run protein over the chip. The protein will only stick to oligos which contain a suitable binding site. Measure fluorescence across chip to determine where proteins were bound. The location on the chip tells you which oligo is at that position. The binding region will only exist in part of the oligo. This region can be predicted using the predetermined binding motif obtained from previous experimentation.
   b) ChIP-seq: Apply formaldehyde to living cells to crosslink all proteins to the DNA it is bound to *in vivo.* Break open cells and sonicate lysate to shear overhanging DNA fragments that aren't protected by the protein of interest. Run product over beads containing antibodies against your protein of interest. Purify DNA from proteins which were bound to the column and sequence. Only sequences which were bound to your protein in vivo will be present.
   c) SELEX-seq is able to measure relative affinity the best because it generates the largest amount of data across several rounds. However, the experiment is performed purely *in vitro* and may not represent the true conditions of the cell. PBM was popular before next generation sequencing became big, but it is quite expensive and does not provide as much information as SELEX-seq today. ChIP-seq is able to capture information about binding under *in vivo* conditions, but you lose a lot of control over which oligos are exposed to your protein of interest. Therefore, you may not be able to identify sequences which would be able to bind tightly, but aren't native to the cell.

3) Completed using R version 3.4.0

4) Prediction models for *in vitro* data

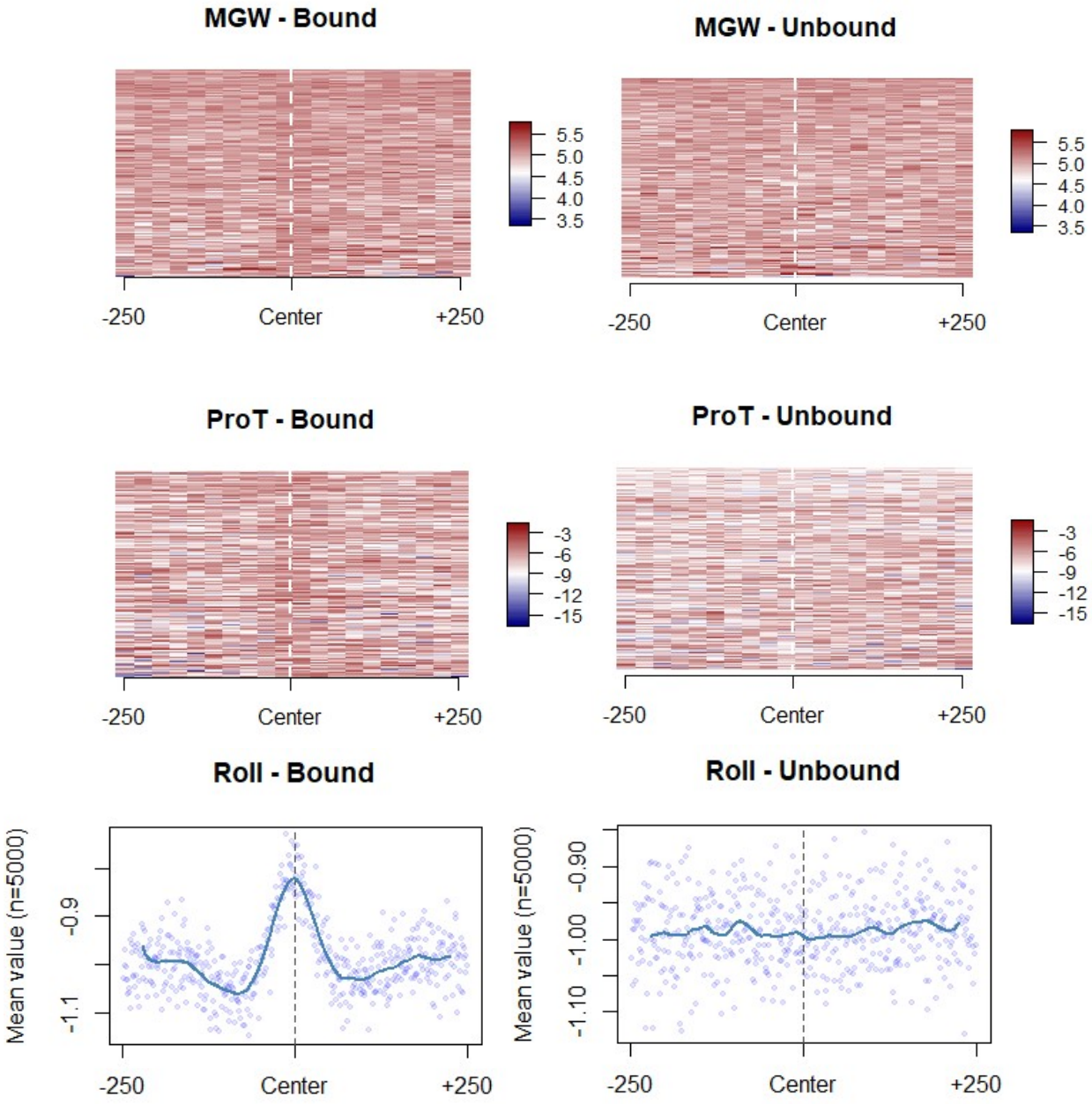
	1-mer (R^2)	**1-mer + shape (R^2)**
Mad	0.7748677	0.8633198
Max	0.7855400	0.8642059
Myc	0.7778829	0.8545234

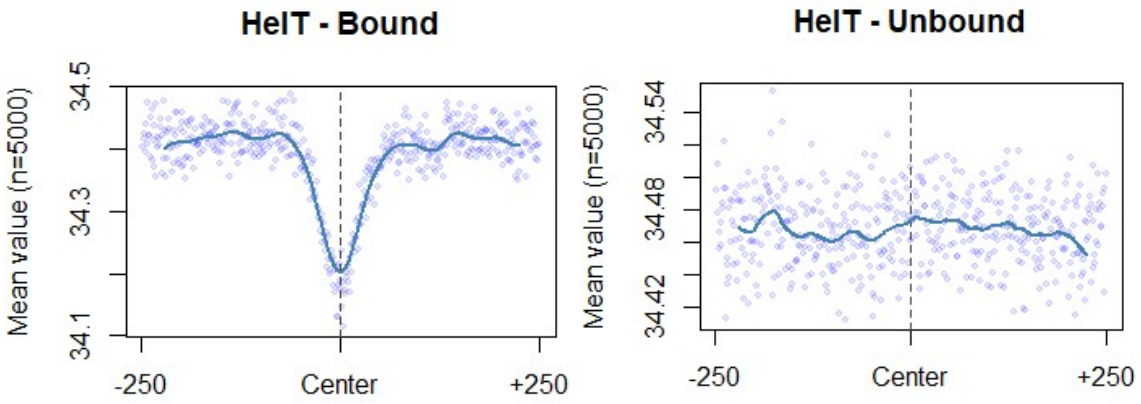
5) Since all points are above the diagonal, the 1-mer + shape model appears to perform consistently better than the 1-mer model at predicting affinity.



6) All files exist and the packages from question 3 are installed.

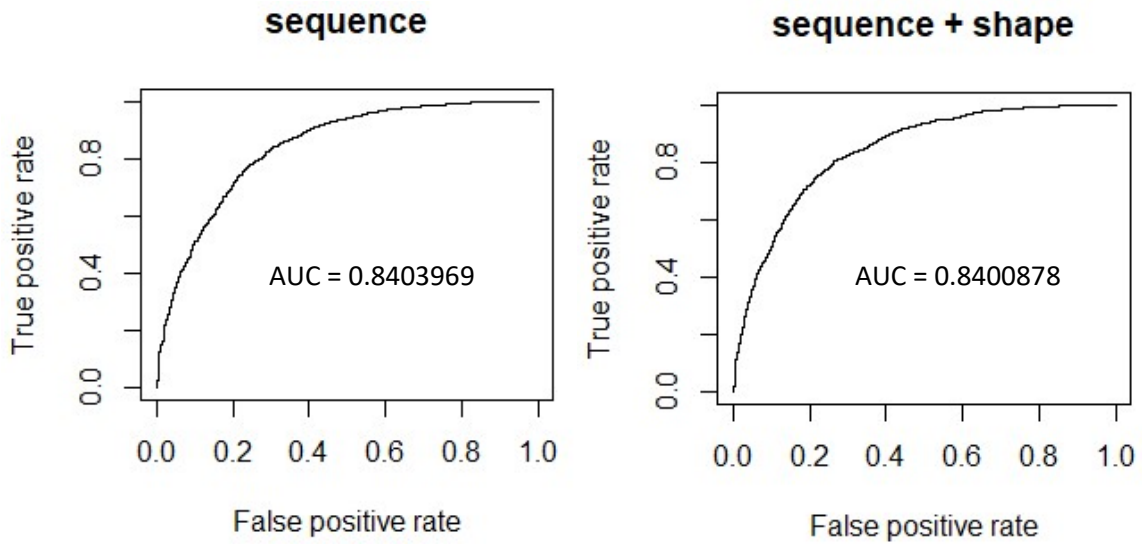
7) High-throughput *in vivo* data analysis





First, it should be noted that in every graph, the deviation from average appears around the center. This deviation is due to the recognition of certain features within the binding site recognized by the protein of interest. The flanking sequence is carried along due to the sonication process. The two heat plots shown suggest that the protein prefers a binding site with a larger than average minor groove width and propeller twist. This is exemplified by the two center columns which appear darker than the rest. The two plot shapes show that the protein prefers a binding site with a larger roll than average, and a smaller helical twist than average.

8) Prediction models for *in vivo* data



Adding shape data did not appear to improve predictive capability of the models based on ChIP-seq data. However, the AUC scored of both models were both quite high in general.