Jordan Bruner
10-25-16

BISC481 Homework 3: Modeling of Protein-DNA Binding Specificity/ Statistical Machine Learning

1. A public repository has been created on my Github account (https://github.com/jbruner/BISC481) and contains a readme. Tsu-Pei has been added as a watcher to the repository. R scripts used in questions 4, 5, 7, and 8 are uploaded.

2.

a) In *in vitro* method SELEX-seq, oligonucleotides are selected by a layer of target proteins. Unbound nucleotides wash away while those that bind to the target proteins are amplified and re-selected for further determination of target protein binding specificity. In *in vitro* PBM, a microarray of dsDNA sequences binds transcription factors that bind fluorescent tagged antibodies, allowing digital imaging of the resulting array to determine the sequences at which the protein bound

b) In *in vivo* ChIP-seq, crosslinking and immunoprecipitation allows accurate sequencing of the DNA of interest

c) While *in vivo* methods may seem to more accurately replicate the conditions of the human body, they are affected by additional factors such as chromatin structure which complicate the specificity of DNA-protein binding. *In vitro* methods allow better control of the testing environment, however results cannot be dependably expected to be the same in living organisms.

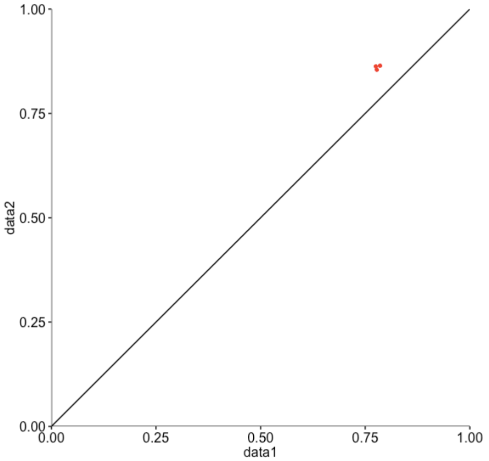
3. gcPBM data was downloaded and all necessary R packages were installed

4. R^2 value of 1-mer sequence model and 1-mer model including shape:

	1-mer	1-mer + shape
Mad	0.77525	0.86268
Max	0.78540	0.86427
Myc	0.77771	0.85456

5.

a) Plot comparing coefficient of determination (r^2) in 1-mer (X-axis) and 1-mer+shape (Y-axis) models:

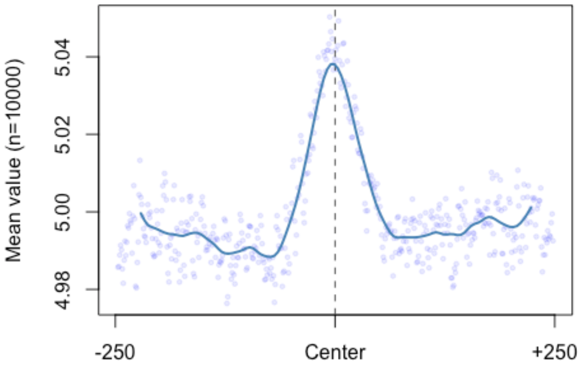


b) This plot clearly indicates that the model generated with the help of DNAshapeR was significantly more accurate in modeling binding specificity than the model considering sequence alone.

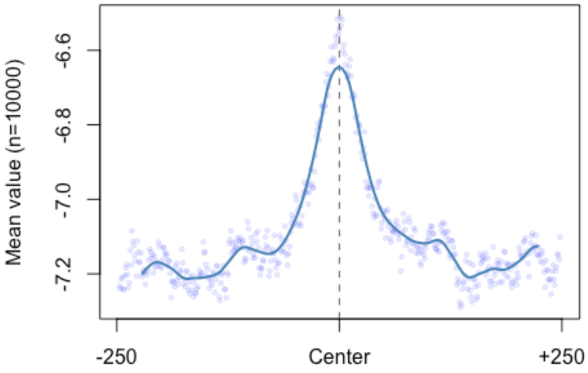
6. ChIP-seq data was downloaded and necessary R packages were installed.

7.

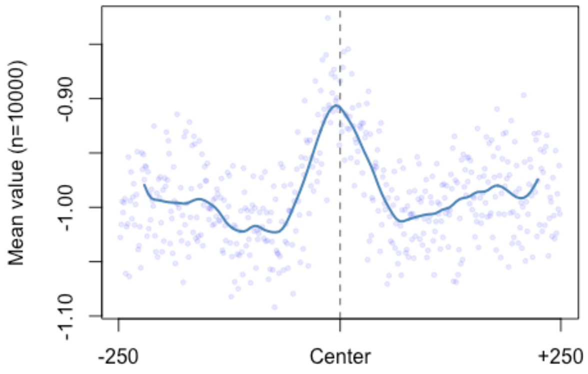
a) Plot of minor groove width along 500bp sequence:



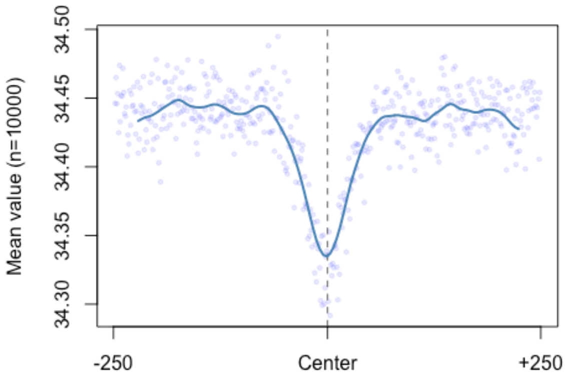
Plot of propeller twist parameter:



Plot of roll parameter:



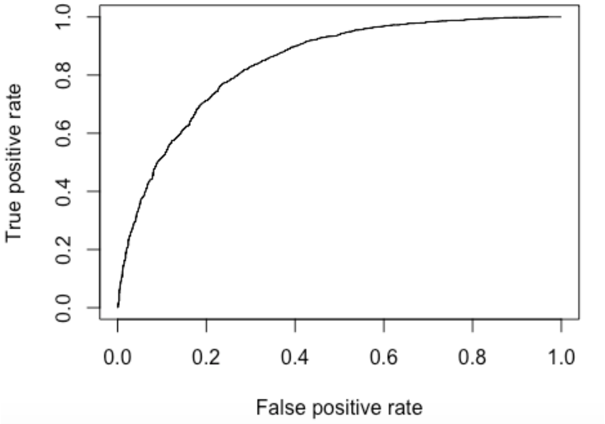
Plot of helical twist parameter:



b) These results consistently show a significant deviation from the typical values for each DNA shape parameter near the center of the 500bp sequence. This indicates that the center of the sequence is likely an important location for DNA-protein binding specificity, as the protein may be able to recognize the altered three-dimensional shape of the DNA macromolecule via the increased minor groove width and other shape parameters.

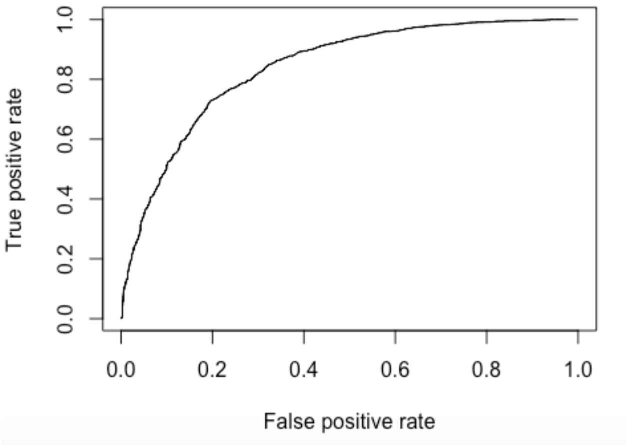
8.

a) Linear regression model without shape ROC:



AUC value: 0.84225

Linear regression model with shape ROC:



AUC value: 0.83968

b) These results indicate that for the CTCF transcription factor, the incorporation of DNA shape in the model has no effect on the accuracy of predictive modeling of binding specificity. Therefore, it can be assumed that CTCF transcription factor is not dependent on shape readout of DNA.