Myrl Marmarelis

Assignment #3

1

`https://github.com/myrlgm/bisc481`

2

(a) *in vitro* SELEX-seq and PBM: the Protein Binding Microarray measures the amount of proteins (tagged with fluorophore antibodies) that have bound to each DNA subsequence which is bound to a slot in the array. SELEX-seq (Systematic Evolution of Ligands by EXponential enrichment) uses this technique at a larger scale in order to generate a matrix of probabilities for the nucleobase at each position.

(b) *in vivo* ChIP-seq uses specific antibodies to separate bound protein + DNA sequences from unbound sequences.

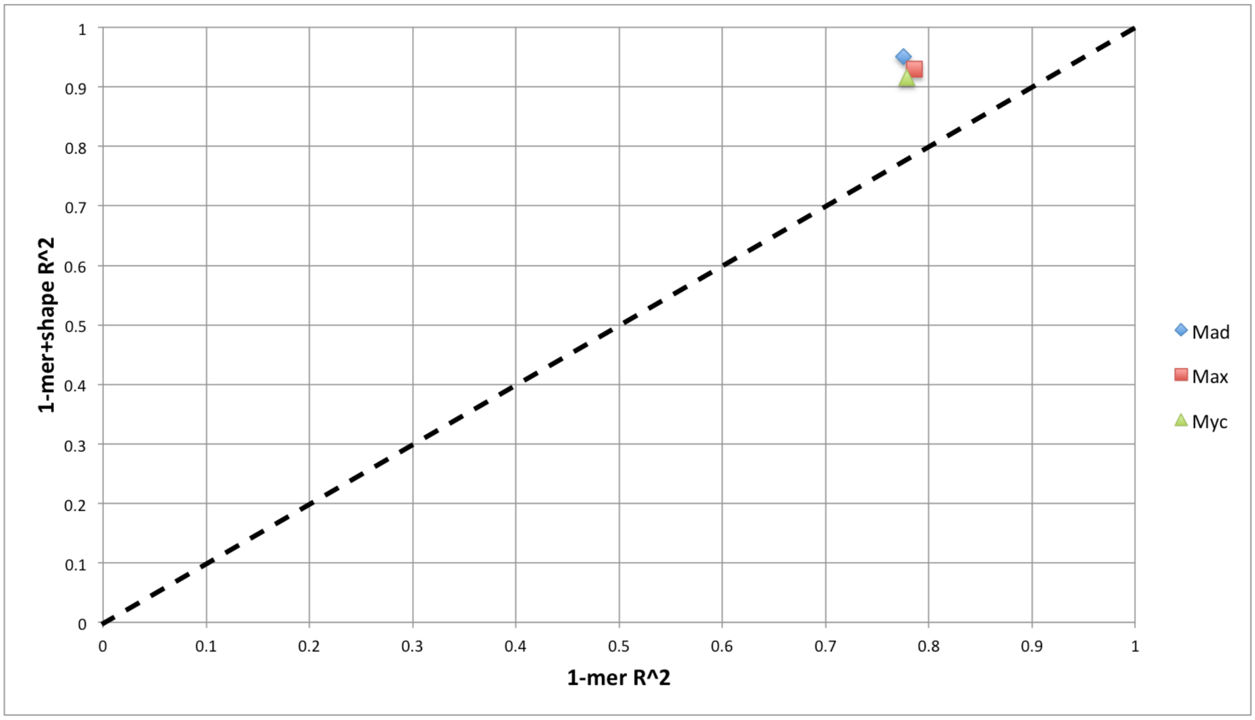
(c) In vitro experiments can give fairly accurate estimates of the actual DNA sequences. In vivo, one can only perform a two-class classification. So there is less information that can be inferred, but we know the environment in the cell is close to that of a living organism, so we can observe different intracellular activities in real time. In vitro, this is not the case because we need to remove the DNA from a cell.

4

1. *Mad*: “1-mer” $R^2 = 0.7754$, “1-mer+shape” $R^2 = 0.9510$.
2. *Max*: “1-mer” $R^2 = 0.7862$, “1-mer+shape” $R^2 = 0.9292$.
3. *Myc*: “1-mer” $R^2 = 0.7787$, “1-mer+shape” $R^2 = 0.9155$.

5

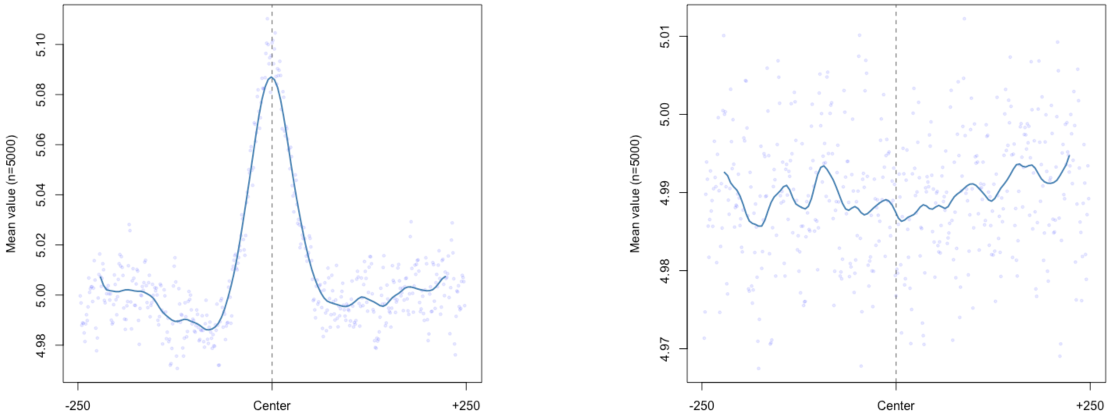
(a)



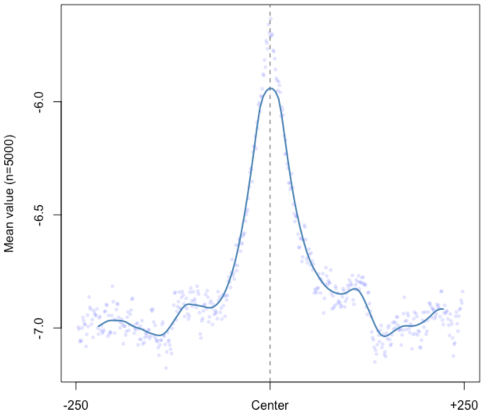
(b) I learned that including shape data in the model's input can greatly increase its accuracy.

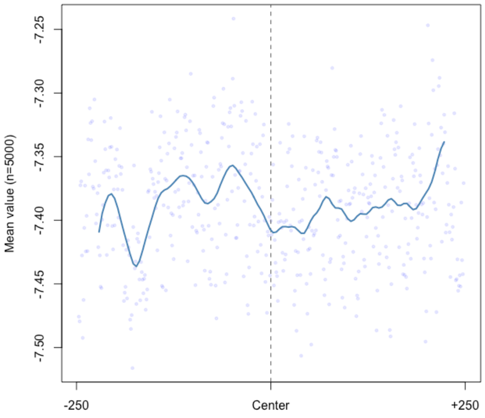
7

(a) minor groove widths for bound and unbound:

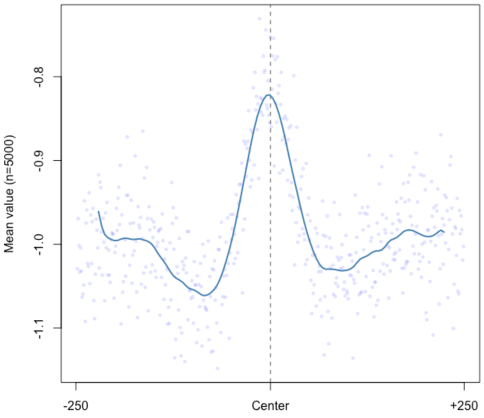


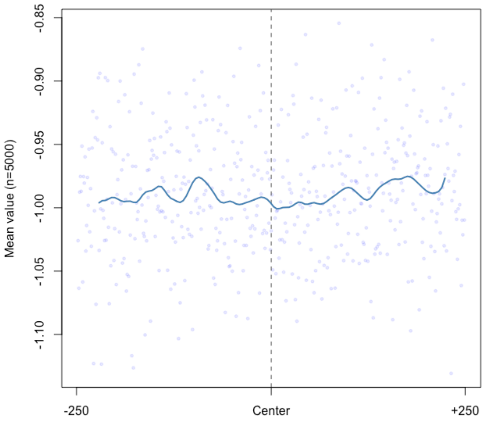
propeller twists for bound and unbound:



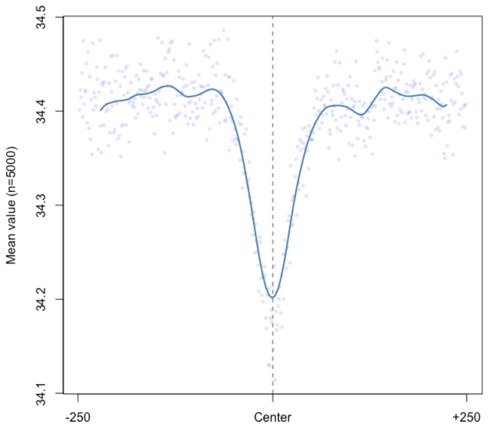


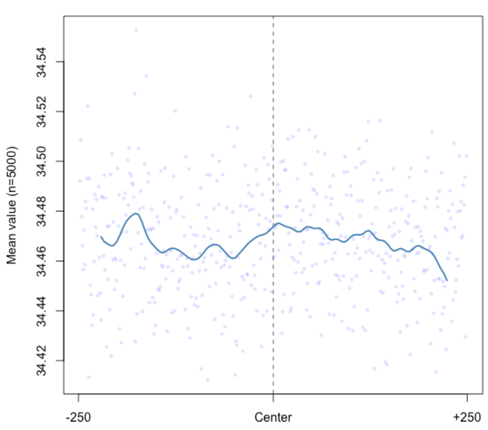
rolls for bound and unbound:





helix twists for bound and unbound:

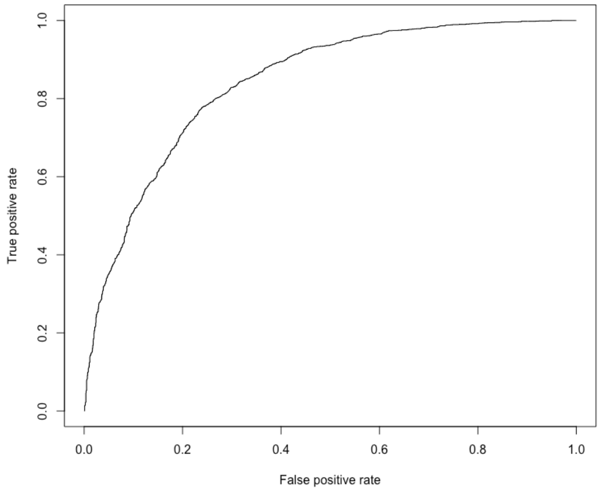


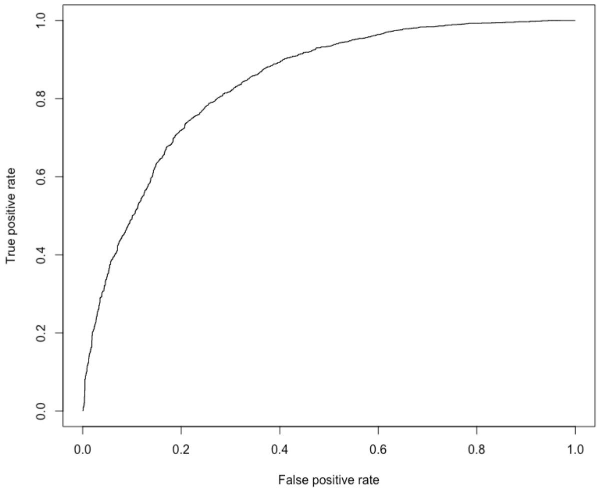


(b) I learned that you can differentiate between DNA shapes by looking at the graphs of their geometrical parameters.

8

(a) ROC curves for the logistic regression models for “1-mer” and “1-mer+shape”:





.

AUC for “1-mer”: 0.8406; AUC for “1-mer+shape”: 0.8398.

(b) I learned that the inclusion of shape data does not improve the performance of the classifier model.