BISC 481, Assignment 3 // Marissa Di

(1) Application of an open-source and distributed revision control project
   a. Repository name: BISC481_marissed
   b. README.md: Marissa Di
   c. Added TsuPeiChiu as collaborator

(2) High-throughput binding assays.
   a. *In vitro*: SELEX-seq and PBM—SELEX-seq, systematic evolution of ligands by exponential enrichment sequencing, is essentially a method where the binding proteins are attached to the plate and the DNA library of binding sites is passed through, and the binding sites that become attached are sequenced to see which binding sites were selected by the protein. PBM, protein binding microarray, is the opposite situation where the DNA is attached to the plate and the proteins are passed through. In both cases the result is quantitative binding data that measures the affinity between the protein and binding sites via signal intensity, and the experiment is done in vitro.
   b. *In vivo*: ChIP-seq—This method combines chromatin immunoprecipitation (ChIP) with sequencing, on a pull-down method that cross-links DNA and protein based on binding. There is a binding peak where the protein binds, and the result is qualitative binding data (binding or non-binding) that is done in vivo. In terms of machine learning, the classification of binding or non-binding is what is predicted.
   c. One advantage of using in vitro analysis methods is that the experiment is relatively uncluttered experiment design; that is, it can provide binding data purely based on the protein-DNA interaction (and nothing else). Also, the information is quantitative, which means that it can provide a spectrum for the binding affinity rather than a black-and-white binding status. However, it requires high read coverage and may not be able to accurately predict binding in a biological context. ChIP-seq, on the other hand, can provide qualitative binding data with a biological basis and categorically determine whether or not a protein-DNA pair will bind with one another. However, it is not able to provide qualitative data with regards to the degree of binding strength, and high sensitivity is expensive due to the sequence tags required, and there may be biological "noise" in the final result.

(3) Preparation of high-throughput in vitro data analysis
   a. Downloaded R 3.3.0
   b. Installed Bioconductor (> source("https://bioconductor.org/biocLite.R") > biocLite())
   c. Installed DNAshapeR package (> biocLite("DNAshapeR"))
   d. Installed the caret package (> install.packages('caret'))
   e. Downloaded the gcPBM data (from GitHub)

(4) Build prediction models for in vitro data:
   a. Used DNAshapeR to generate feature vectors for "1-mer" sequence and "1-mer+shape" models for Mad, Max, and Myc. Based on MLR example done in class, did "P4_...seq.R" and "P4_...seq-shape.R" files to generate sequence and shape information (feature

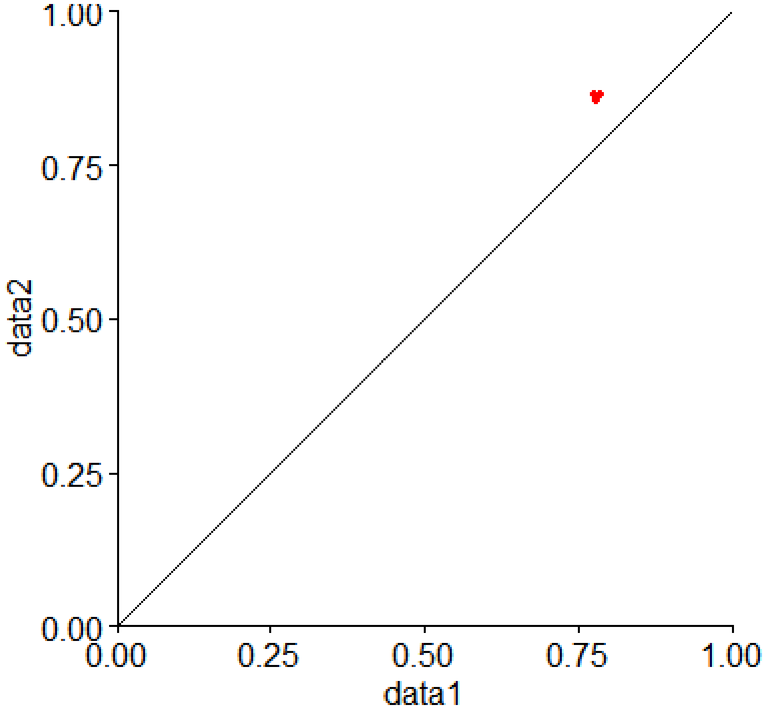
vectors) for the 3 datasets based on their associated fasta files. Refer to "P4_gcPBM_MLR.R" for changed lines (23, 27).

b. Used caret, built L2-regularized MLR models for "1-mer" and "1-mer+shape" features with 10-fold cross validation. Average R^2:

	Mad	Max	Myc
Sequence	0.775516	0.785694	0.778183
Sequence + Shape	0.863464	0.864357	0.854206

(5) High-throughput in vitro data analysis

a. Plot to compare 1mer vs. 1mer+shape models



Used "P5_SEQ-SHAPE_plot.R" to plot the data. Data1 is the 1mer sequence-only model R^2 values for Mad, Max, and Myc, and Data2 is the 1mer+shape sequence+shape model R^2 values for the same data sets, plotted against each other.

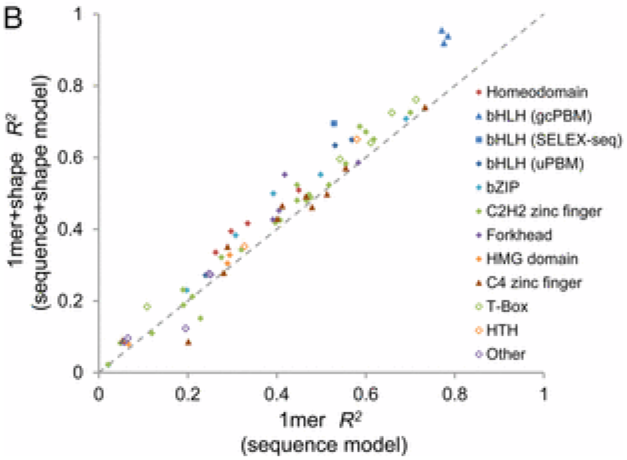


Figure 1B of Zhou et al. PNAS 2015

Same axes as the plot generated above, included for reference.

b. The results show that for the PBM data from this experiment, the 1-mer+shape model is slightly but consistently better than the 1-mer model, according to R^2 values (~0.86 vs. ~0.78, respectively). The R^2 value summarizes how well the models correlated with the actual data, and for all 3 datasets Mad/Max/Myc the 1-mer+shape model had higher correlation than for the 1-mer model. This means that including shape information in the model can significantly increase its accuracy for these in vitro data. Even though the models are both created with MLR, the inclusion of the extra information makes one of the models better than the other, which means that for these particular proteins, the additional shape data is particularly important to the binding predictions. However, $R^2 \approx$ 0.8 from both models is fairly good nonetheless (the closer to 1, the better).

   The 10-fold cross validation allows for the whole datasets split into parts to be used for both training and testing, so that the model is not unfairly weighted towards one portion of the data. The L2 regularization also prevents the model from including unnecessary parameters (or at least decreasing their relative weight in the model) and prevents overfitting.

(6) Preparation of high-throughput in vivo data analysis

a. Downloaded M. musculus ChIP-seq CTCF data, bound and unbound

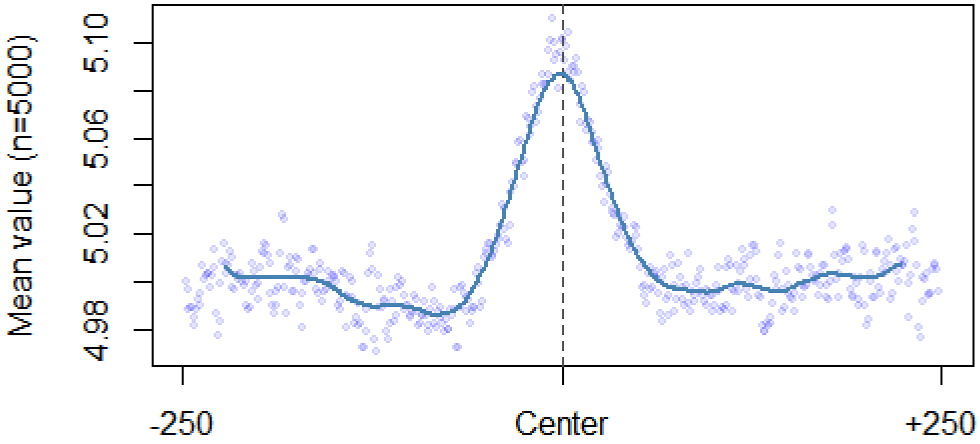
b. R packages installed

(7) High-throughput in vivo data analysis

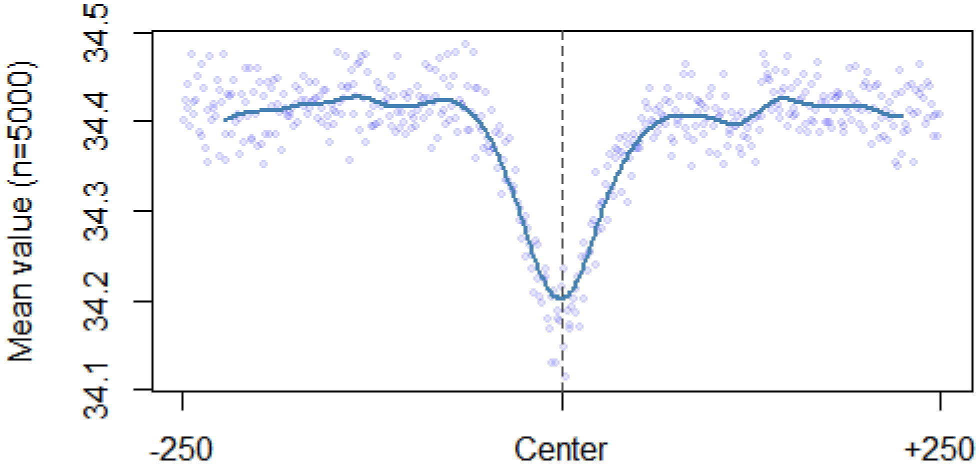
a. Used plotShape() to generate ensemble plots for the DNA shape parameters of minor groove width (MGW), propeller twist (ProT), Roll, and helix twist (HelT). R scripts used are “P7_CTCF_ensemble_BOUND.R” and “P7_CTCF_ensemble_UNBOUND.R”. Refer to “P7_CTCF_ensemble.R” for changed line (13).

BOUND_500 PLOTS

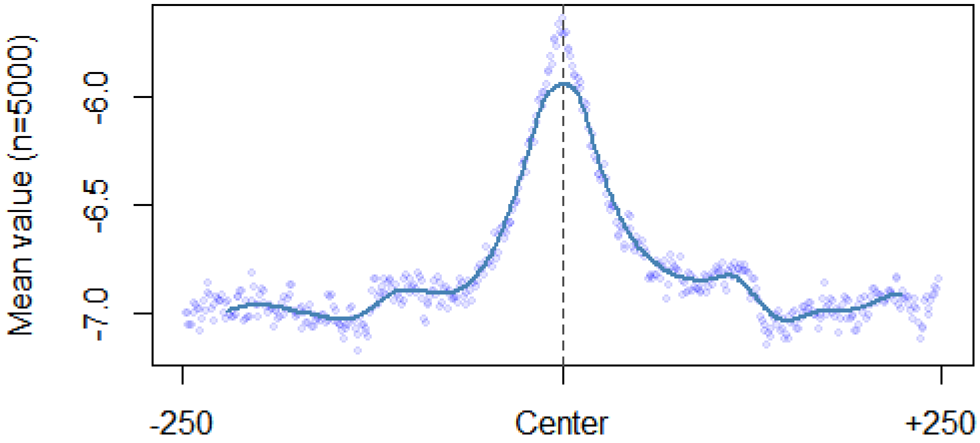
Minor Groove Width



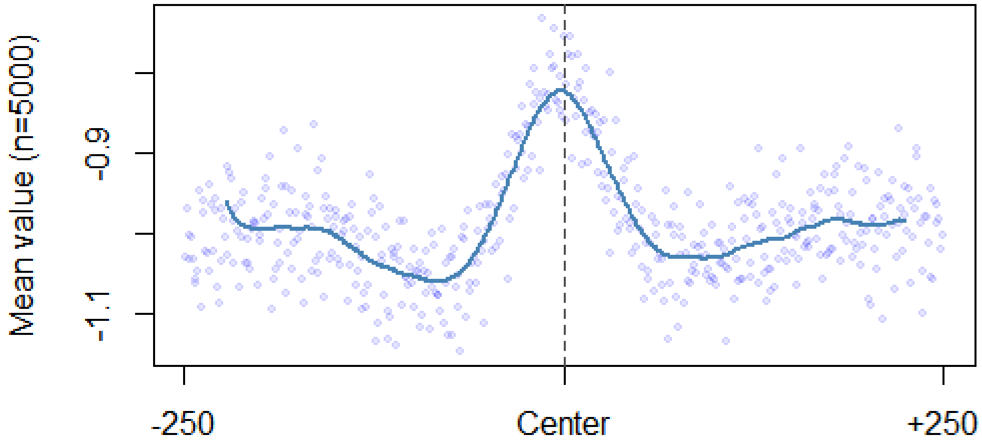
Helix Twist



Propeller Twist

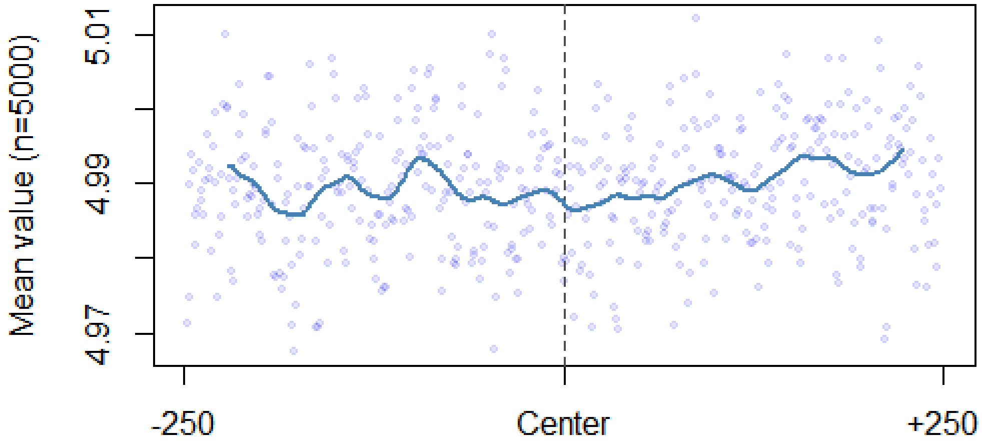


Roll

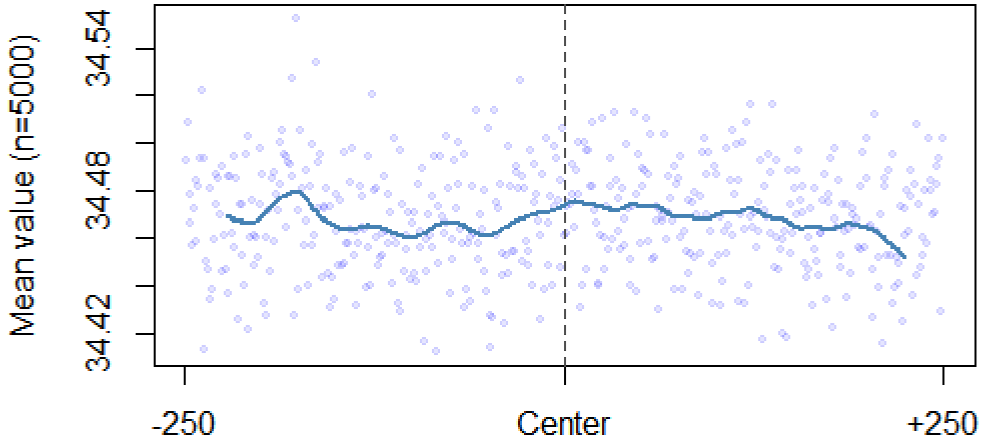


UNBOUND_500 PLOTS

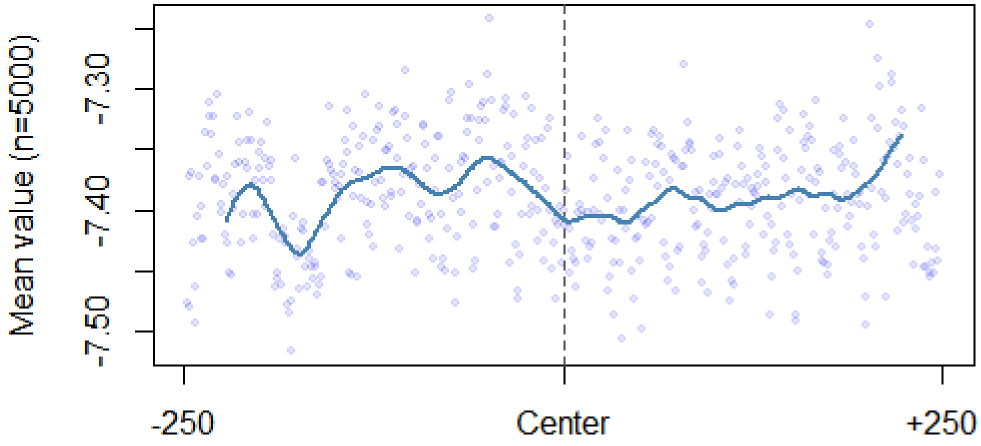
Minor Groove Width



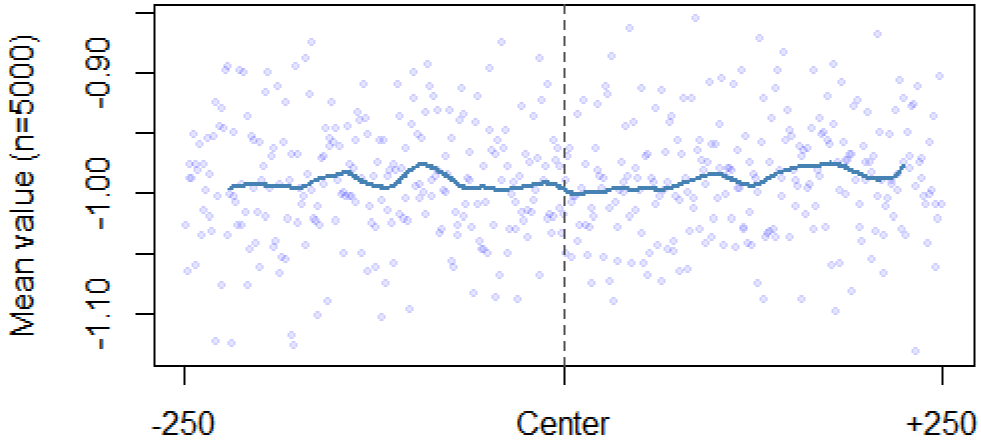
Helix Twist



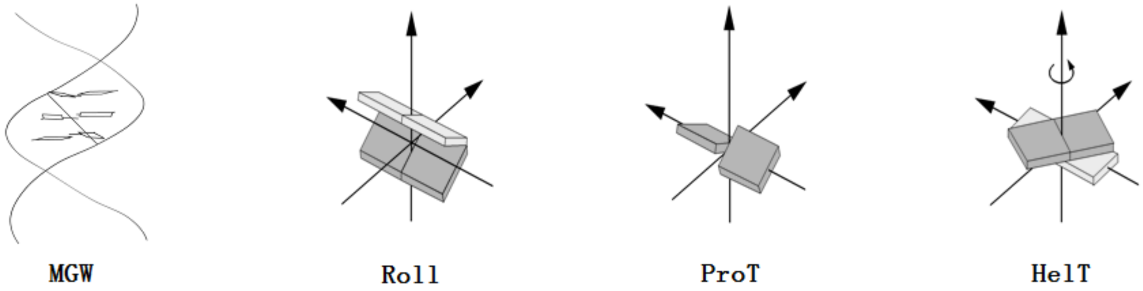
Propeller Twist



Roll



b. From the results, a clear difference between the different structural parameters for the CTCF binding between bound and unbound sequences can be seen. Where the bound data has a clear difference in the data (either a spike or a decrease) at and near the binding site, the unbound data does not have any significant difference in features near the center of the sequences; all of the unbound plots have roughly equal values for the parameters regardless of sequence position, and their individual plots have much more scattered data points than any of the bound plots.

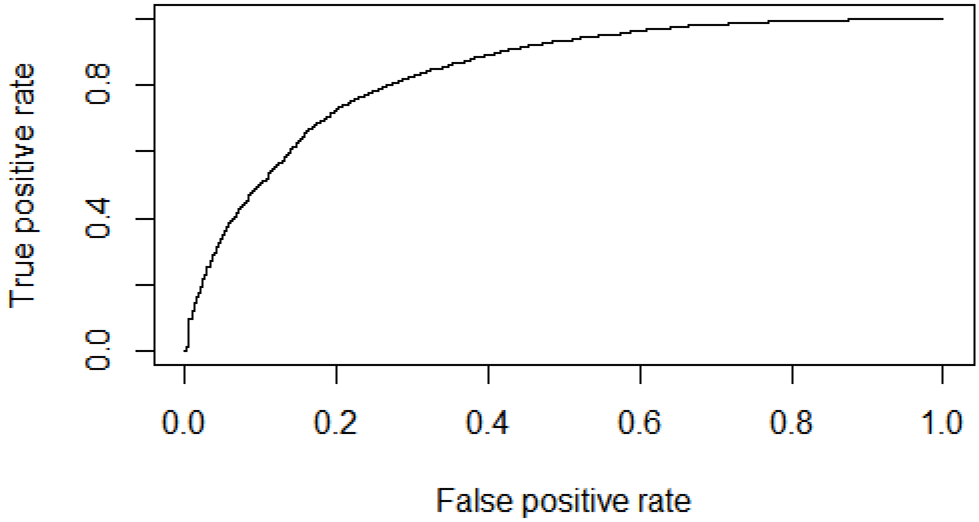


The bound sequences have structural features that presumably allow the protein to recognize and bind to the sequence: the minor groove width, propeller twist, and roll are increased, and helix twist is decreased. It is possible that—similarly to how a narrow minor groove can increase electrostatic potential and attract arginine residues more easily—the increased minor groove width changes the local electrostatic potential and affect binding specificity. Propeller twist is an intra-base pair parameter, and the helix twist and roll are inter-base pair parameters. The combination of increases and decreases means that at the center, the bound sequence is more "unwound" than the rest of the sequence (decreased helix twist) and more "stretched out" along its base pair sequence (increased roll and propeller twist), which could allow for easier access of the protein to the sequence and increased space for the protein to bind, allowing the protein to recognize the binding site because of its shape difference from the surrounding sequence.

(8) Build prediction models for in vitro data

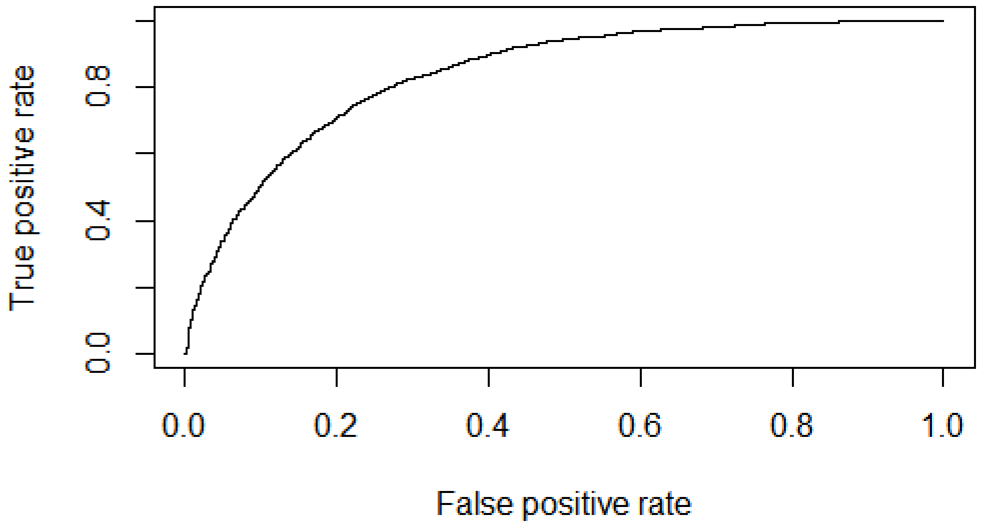
a. Logistic regression models for "1-mer" and "1-mer+shape" features, plots of ROC curves and AUC scores below. Used "P8_logreg_seq.R" and "P8_logreg_seq-shape.R" for 1-mer and 1-mer+shape curves and scores, respectively. Refer to "P8_logreg.R" for changed line (46).

1-mer+shape



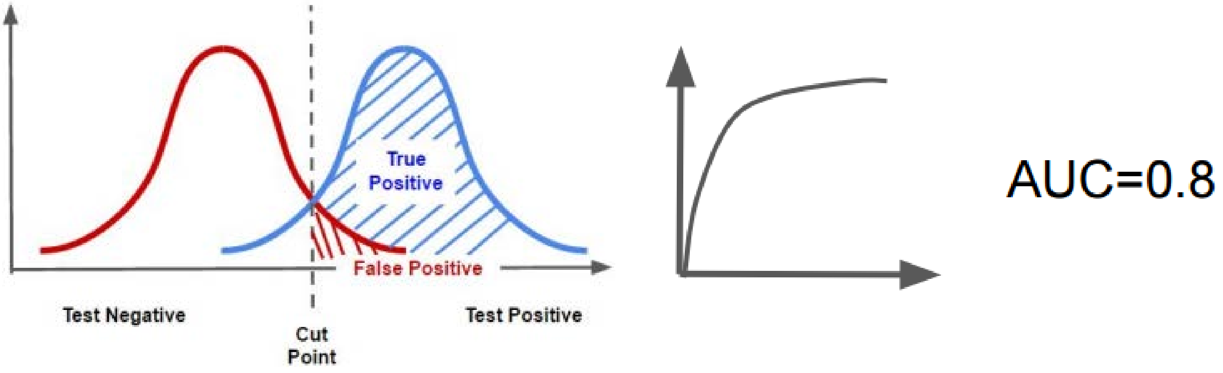
AUC = 0.8403172

1-mer



AUC = 0.8402609

b. Looking at these plots, it can be seen that the two models, 1-mer and 1-mer+shape, are quite similar in terms of their ability to predict true positives vs. false positives. The AUC score for the 1-mer+shape is marginally better than the 1-mer model score (the closer to 1, the better).



However, both are relatively "good" models with scores of 0.84 as to how well they can predict bound vs. unbound based on sequence, at least according to the training and testing data. So for CTCF binding data obtained from in vitro methods (SELEX-seq, PBM), the logistic regression is also a valid and relatively accurate method of predicting binding status. Also, between the 1-mer and 1-mer+shape models, there is not one that is significantly better than the other (based on AUC score), as opposed to the MLR when the 1-mer+shape was consistently slightly better than the 1-mer (based on R^2 value). The reason that they are quite similar might be because the shape data of the protein does not contribute much to the accuracy of the model, so the shape of the protein (independent of sequence) is not as important to predict the binding. In fact, adding the shape data might in some cases cause an overfitting problem and *decrease* the accuracy of the model, although in this case it seems that there is no significant disadvantage to including shape data.