Alyssa Sanchez

BISC481, Fall 2016

Third Assignment for 1st Section of the Course (Rohs)

Modeling of protein-DNA binding specificity/Statistical machine learning

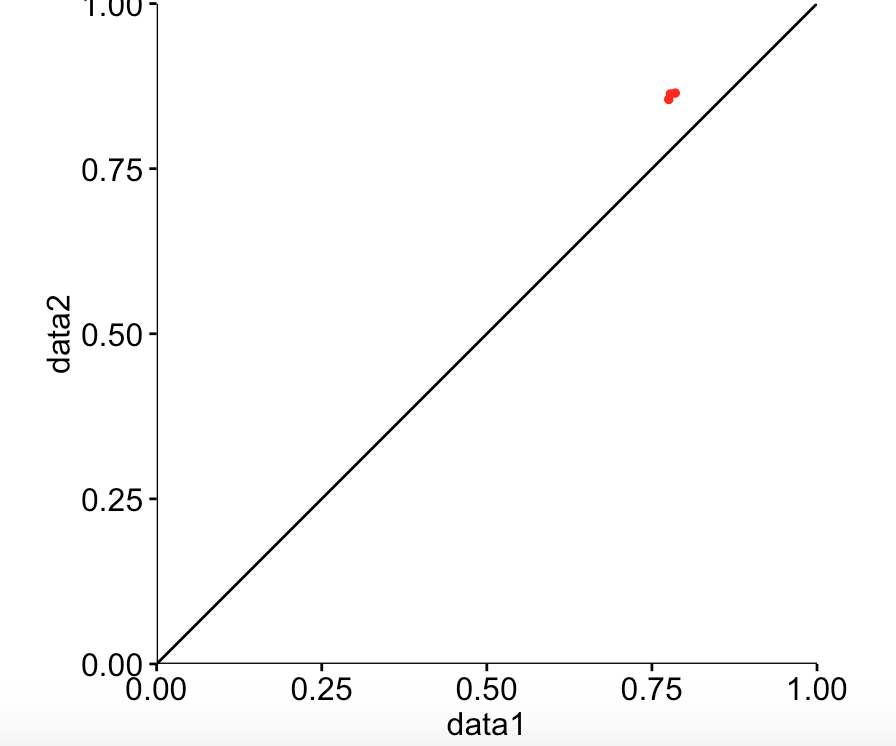
1. a. Repository created on GitHub.

b. Name ‘Alyssa Sanchez’ is in the file of README.md

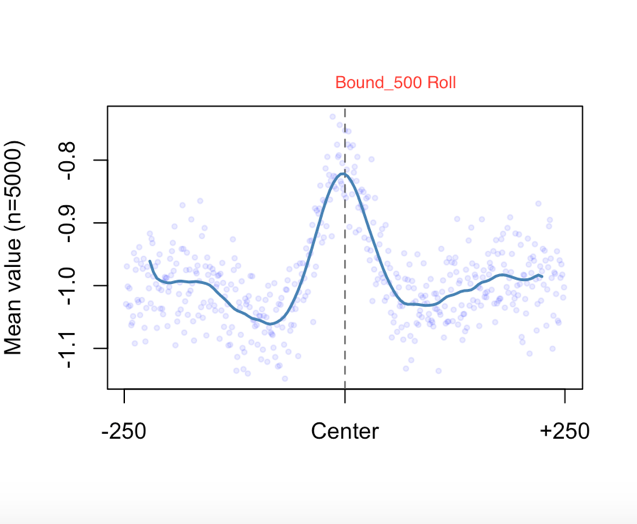
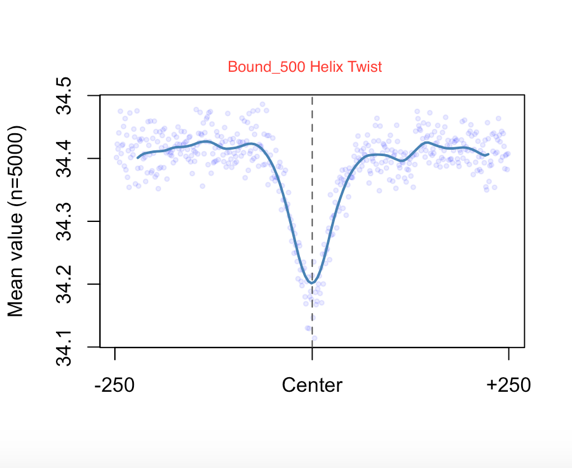
c. Collaborators TsuPeiChiu added

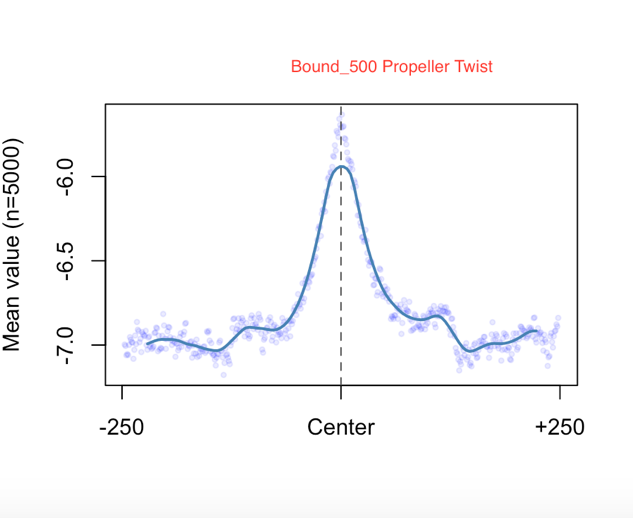
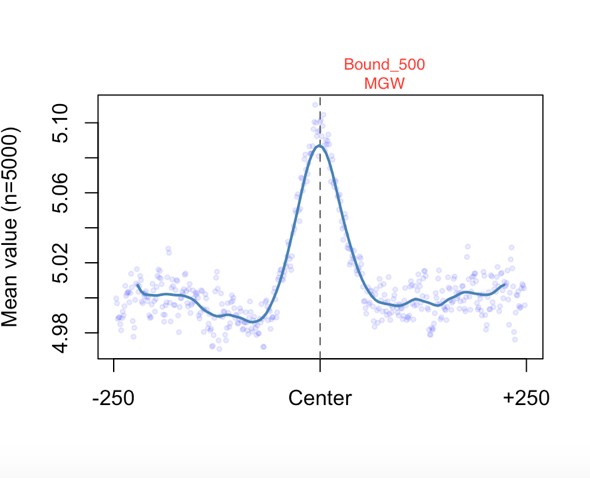
1. a. Protein Binding Microarrays (PBMs) use a double stranded microarray to bind with a epitope tagged transcription factor to measure the fluorescence of epitope tagged proteins bound to their binding sites. The advantages of PBMs are that they are inexpensive, fast, and provide relatively accurate measurements of binding. Disadvantages of PBMs include that these microarrays only encompass 10 binding sites. This makes it hard to model binding sites that have more than 10 base pairs. The SELEX-seq method uses parallel sequencing with protein DNA SELEX assays. A nucleotide that has a random regions that is defined by primer sites is used to bind the protein of interest. The DNA bounded by the protein is separated by the DNA that is unbounded and is amplified with PCR. Advantages include that there is no limit to the size of the binding site, and the binding data of large protein complexes can be captures. Disadvantages include that this technology is expensive and that there can be strong sequence faults in the many rounds of PCR. ChIP-seq uses chromatin immunoprecipitation along with DNA sequencing to identify protein-DNA interactions and binding sites. Advantages of ChIP-seq are that there are opportunites for high resolution and reduction in noise of data. The biggest disadvantage of ChIP-seq is that it is very expensive.
2. All software was installed in order to obtain the R squared values of 1-mer and 1-mer+shape for all three mad,max, and myc data sets.

|  |  |  |
| --- | --- | --- |
|  | 1-mer | 1-mer+shape |
| Mad | 0.7754663 | 0.8631336 |
| max | 0.7853483 | 0.8644654 |
| myc | 0.7779299 | 0.854758 |

1. 

The plot above is indicative of a comparison of the two different models (1-mer and 1-mer+shape) of the three different data sets. As you can see, each R-squared point is located within the same region. This means that our data is positively correlative and both the 1-mer and 1-mer+shape of the three data sets (mad, max, and myc) have a correlation of the sequencing technique used. Using the DNA shaper with R studio, we are able to compare the sequence and shape model (data 2) with the sequence model for all the three data sets. This way, we are able to predict the DNA shape with this function and the calculation of R-squared.

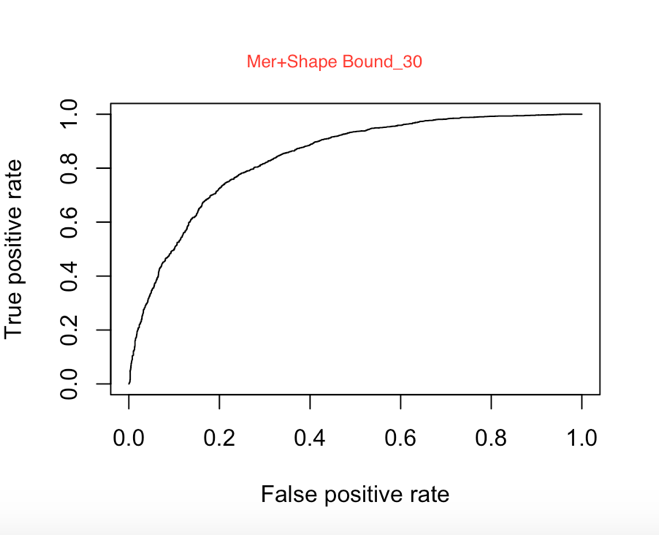
1. The bound and non-bound data of the CTCF transcription factor was downloaded.
2. A.



The plots above represent the roll, twist, helix twist, and minor groove width within the CTCF transcription factor. Looking at the minor groove width, we can say that the active site of this protein is within the center region. We also know that the propeller twist and the roll parameters are within this same site but we do know that there is not much of a helix twist within this same region of the protein. Because all of these parameters are important for protein-DNA binding interactions, they are used to predict these binding sites.

1. AUC Mer+shape: 0.8399342

AUC Mer: 0.8418342

The plots below are the ROC curves for the 1-mer and 1-mer+shape features. We know that if the AUC is closest to one, there is no false positive. Since the AUC values were in the 0.8 range, there is a slight false positive in our data between both the Mer+shape and 1-mer for bound 30 CTCF transcription factor although we can say this is generally good data and that the protein and DNA are bound together. The accuracy of the CTCF protein binding to the DNA is shown by these plots and by the AUC calculations. T