Farah Najib

BISC 481

T,Th 11:00 am

10/25/16

Third Assignment

1. See github repository. Both Beibei and Tsu-Pei are collaborators.
2. In vitro and in vivo experiments
   1. SELEX-seq and PBM
      1. SELEX-seq: It is a high-throughput technique for characterizing DNA-binding specificities of transcription factors. Proteins are attached to a surface (a “plate”) and then a flow of DNA sequences from a library is passed through the plate with the attached proteins. The proteins will bind to certain DNA sequences. This can be done for many rounds, which helps with enrichment. You can learn what DNA sequence will be bound to what protein on the plate. You can count how often certain sequences are bound to certain proteins. This gives binding strength between DNA sequences and proteins.
      2. PBM: It is a method that allows for rapid, high-throughput characterization of the in vitro DNA binding-site sequence specificities of transcription factors. It consists of a support surface such as a glass slide, nitrocellulose membrane, bead, or microtitre plate, to which an array of capture proteins is bound. Probe molecules (e.g. DNA fragments of varying lengths), typically labeled with a fluorescent dye, are added to the array. Any reaction between the probe and the immobilized protein emits a fluorescent signal that is read by a laser scanner

Its main advantage lies in the fact that large numbers of proteins can be tracked in parallel.

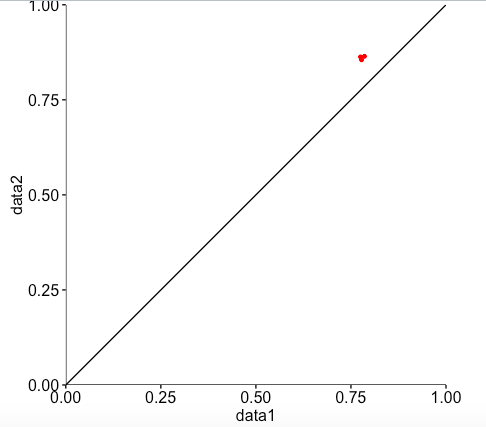
* 1. ChIP-seq: It is a method for identifying genome-wide DNA binding sites for transcription factors and other proteins. It allows for analysis of protein interactions with DNA. The ChIP process enriches specific crosslinked DNA-protein complexes using an antibody against the protein of interest. all the resulting ChIP-DNA fragments are sequenced simultaneously using a genome sequencer.
  2. Advantages and disadvantages: The advantage of SELEX-seq is that it is a method that can give the binding strength between DNA and proteins. You can get quantitative binding data on affinity. PBM is advantageous in that it can track large numbers of proteins in parallel. As with SELEX-seq, it gives quantitative data on binding affinity. ChIP-seq is beneficial it can selectively enrich for DNA sequences bound by a particular protein, but the disadvantage to it is that it only gives qualitative binding data, such as whether or not a protein can bind to a DNA sequence, and not how strong this binding may be.

1. Did this in my computer to answer the other questions.
2. Average R2 for 1-mer sequence model and for 1-mer + model

|  |  |  |
| --- | --- | --- |
|  | 1 mer | 1 mer+shape |
| Mad | 0.774981 | 0.863 |
| Myc | 0.7778097 | 0.8551015 |
| Max | 0.7854914 | 0.8644255 |

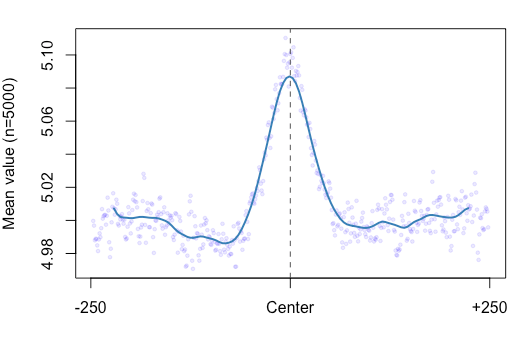
1. Comparison of 1mer and 1mer+shape
   1. Plot

Note: “data1” is 1mer and “data2” is 1mer+shape

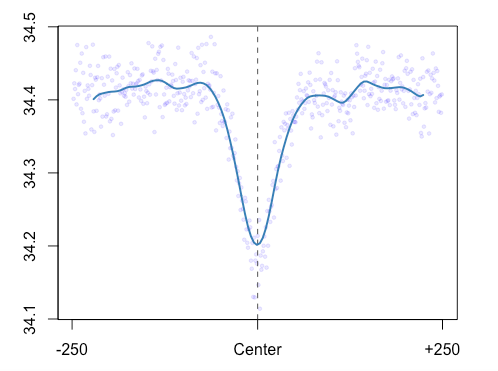


* 1. The two models 1mer and 1mer+shape show results that are very close to each other. The R2 values for both with respect to Mad, Max, and Myc are all very high (around 0.7 – 0.86). This tells us that it is likely that both shape and sequence help the prediction.

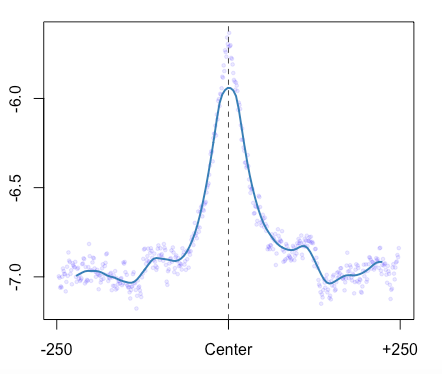
1. Did this in my computer to answer the other questions.
2. Plots for DNA shape parameters
   1. Minor Groove Width (MGW)



Helix Twist (HelT)

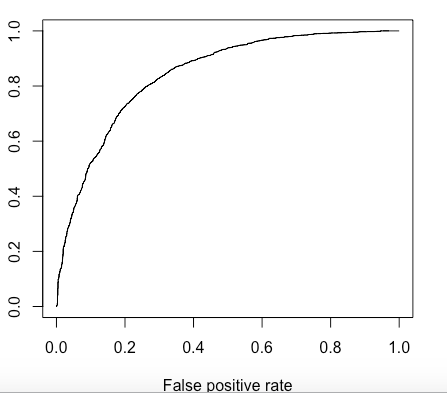


Propeller Twist (ProT)

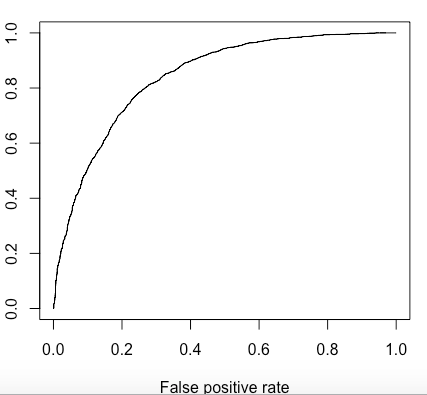


* 1. You can use these DNA shape parameter plots to predict the shape of the DNA. We learn that by using the sequence and running the sequence through the R program that we generated to specifically predict shape, we can predict the shape of the DNA, as can be seen here in these plots that show the different aspects of the DNA in question. The minor groove width is very large at the center. The lowest point of the helix twist is at the center. Also, the propeller twist seems to be much less negative at the center than everywhere else.

1. Prediction models for in vitro data
   1. ROC curve for 1mer (Note: y-axis should say “True positive rate”. Also, AUC score is 0.841013)



ROC curve for 1mer+shape (Note: y-axis should say “True positive rate”. Also, AUC score is 0.8420485)



* 1. The ROC curves give us a means to validate the prediction result. Both curves show that the results are more accurate the closer the curve follows the left-hand border and top border of the graph. For both curves above, the calculated AUC scores are high (~0.84 for both curves), which says that these are good tests for the in vitro data.