BISC 577a, Part3, 2nd assignment

Yingfei Wang, 2319220751

(2)

1. PBM: in vitro experiment, dsDNA with different sequences are put into a microarray, then the protein in the study (for example, the TF of interest) is going to be flowed through the microarray, so that it will bind to its binding sequence (which appears in some of the dsDNA probes in the microarray). After that, a fluorescent label is flowed through the microarray chip to bind the protein in study in order to identify its binding place. By scanning the microarray, we could know the physical location of the fluorescent and thus the binding sequence of the protein in study.

2. SELEX-Seq: in vitro experiment. SELEX-Seq fixs the protein of interest to a platform and flow the dsDNA (a random pool of olignucleotides) through the platform. By doing so, the target sequence of the protein will be bounded by the protein and thus stay in the platform. After flowing the oligonucleotide, the dsDNA fragements are released from the binding protein and go through sequencing. After that, we can know the binding sequence of the protein of interest.

3. ChIP-Seq: in vivo experiment. The experiment is done in cells. First of all, the proteins are cross linked to their binding DNA, and then the cells are lysis, the DNA are sonicated into fragments (while their binding proteins are on them). Secondly, the protein of interest are selected from the pool by using its specific anti-body (this step is called immune-precipitation). By doing so, the target DNA sequence bound by the protein of interest is selected from the pool of DNA fragments. After immune-precipitation, the fragments go through reverse cross link to retain only the DNA. The final step is to sequence the DNA to get the binding motif of the protein of interest. (if the final step is done by microarray, then this experiment is usually referred to as ChIP-chip)

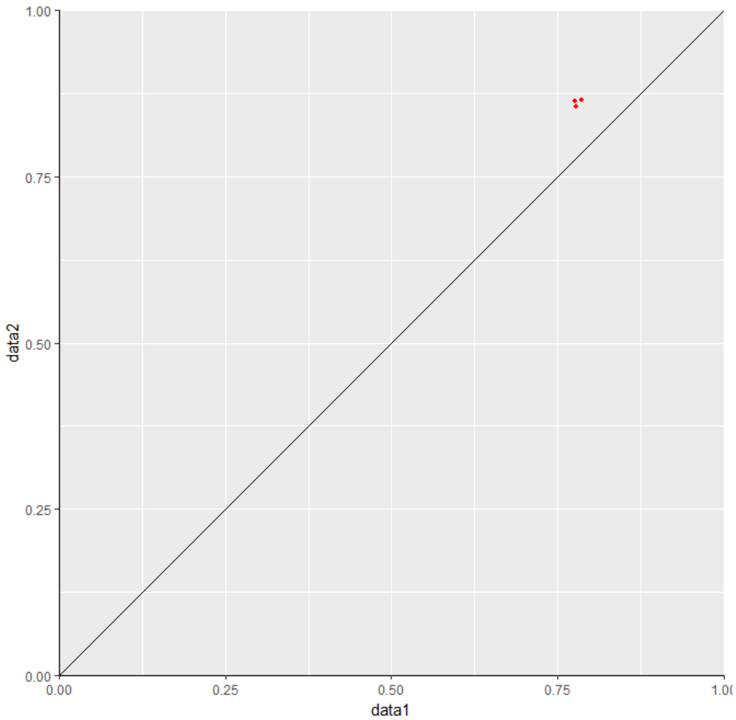
Comparison of experiments: first of all, in vivo experiments are usually more desired than in vitro study, since it better recapitulate the real situation in cells. But in vitro study will be much faster and have higher throughput than in vivo experiments. In addition, the accuracy and correctness of the in vivo study can sometimes be confounded by the person who carried out the experiment, while from this perspective the in intro study is more reproducible and less error-prone.

(4)

the two feature vectors are created using featureType <- c("1-mer", "1-shape") and featureType <- c("1-mer"). The results (Rsquare) are summarized in following table:

Data	1-mer	1-mer+shape
Mad	0.7751104	0.8631741
Max	0.7855506	0.8644418
Myc	0.7776267	0.8550499

(5)



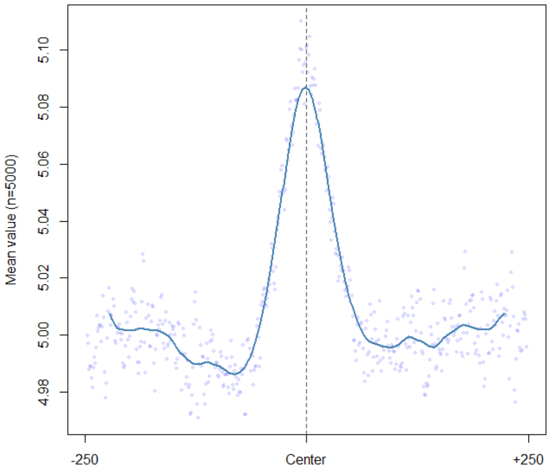
We can see from the graph that all three points are above the diagonal line, indicating that the incorporation in model of shape feature increase the prediction accuracy

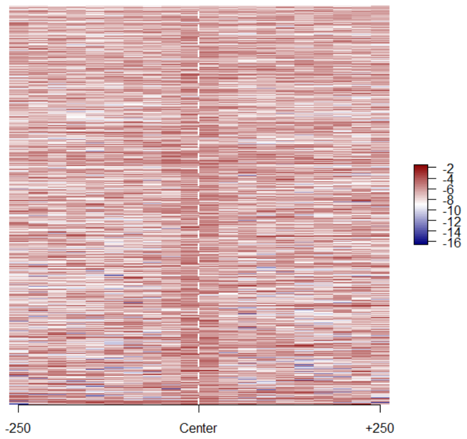
(7)

For bound 500:

MGW

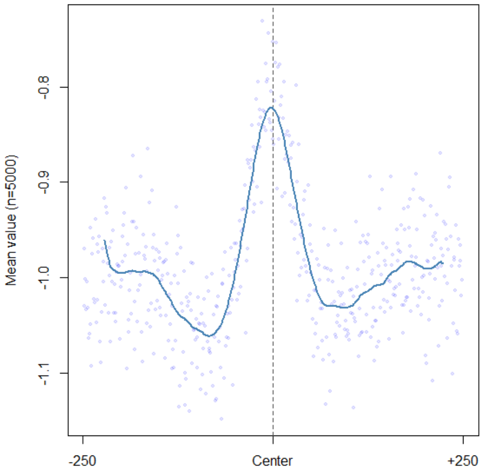
ProT

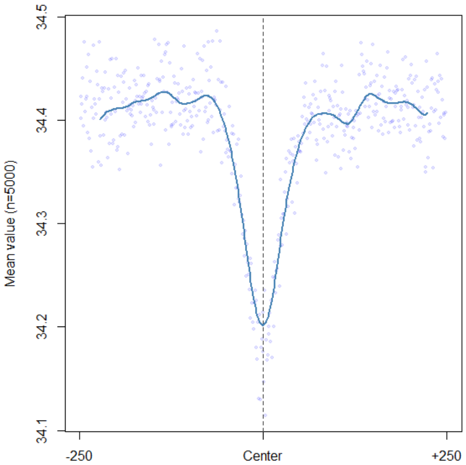




Roll

Helt



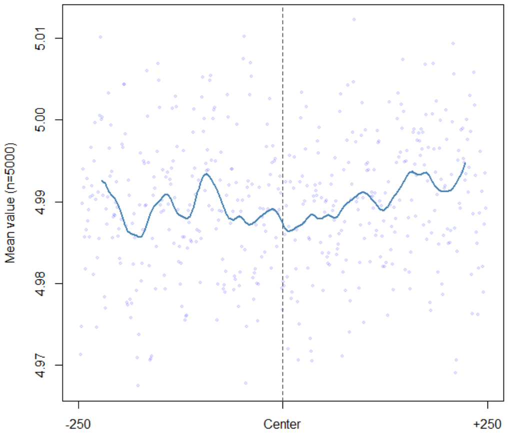


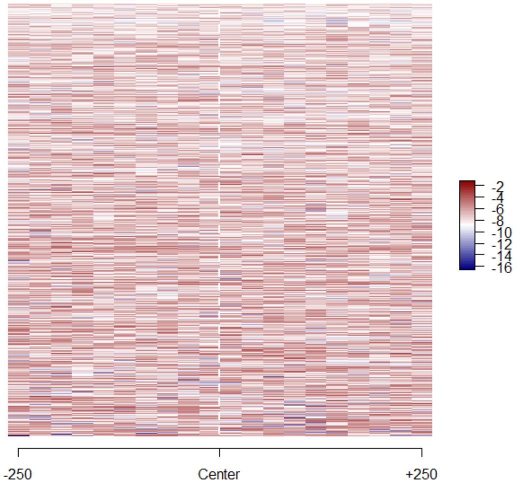
(8)

For unbound 500:

MGW

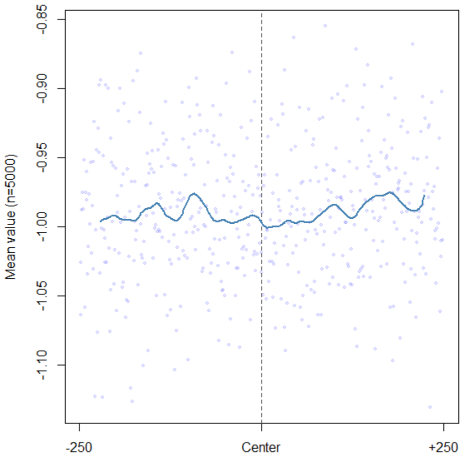
ProT

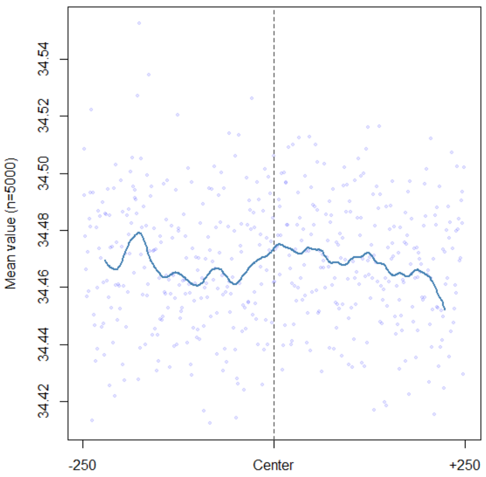




Roll

HelT



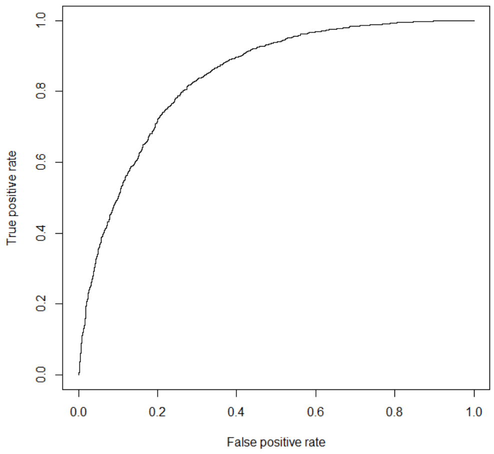


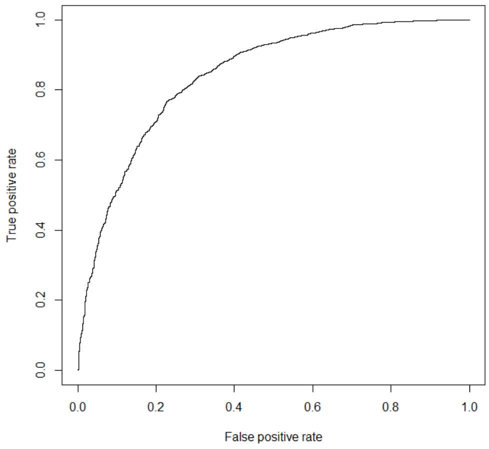
Observations: big difference can be observed between bound sequences and unbound sequences. For example, when comparing the MGW, Roll and HelT, bound motif/sequences shows obvious patterns in the middle of the sequence, while unbound motif shows no concord patterns. In addition, when looking at the ProT, we can see bound sequences have comparatively more positive propeller twist when compared to unbound sequences, also, in the middle of the sequence, a strong positive propeller twist can be seen in the bound sequences.

These differences are of no surprise to us. Since the TF only bind to certain areas in genome although the same motif sequence appears throughout the DNA (since motifs are usually short and degenerate, rendering the likelihood of multiple copies occurring throughout the genome due to tandem chance as high). As we can see, structural characteristics play an important role is terms of TF binding, since the bound sequences show very clear structural difference and unique but concord patterns, when compared with unbound ones.

(8)

Left: 1-mer model, AUC=0.8415835 , Right: 1-mer+shape model, AUC=0.8411131





Observation: adding the shape attributes can’t help distinguishing bound motifs from unbound motifs, although this seems counterintuitive with what we observe in Q6.