BISC577A Unit3 – Assignment 2

Shuo Li 2853594360

(2)

(a) Both PBM and SELEX-seq are in vitro experiments quantifying the interaction between transcription factors(TFs) and their DNA binding sites. PBM contains all possible sequences of length k. Then they bind epitope-tagged TFs to the microarray, and label the TFs with fluorophore-tagged antibody to epitope. Finally, they scan the microarray and decide the TF binding site and affinity. SELEX-seq starts from a pool of synthesized DNA oligonucleotides containing a region of random base pairs. The pool is further sequenced, and exposed to TFs. Then DNAs bounded to the complex of interest are isolated and amplified by PCR, which form an enriched pool. Then affinity-based selection is repeated multiple times. Finally, they integrate the data from multiple runs, and decide the DNA binding site and affinity.

(b) ChIP-seq is an in vivo experiment discovering protein binding site. The procedure of ChIP-seq experiment is first to cross-link protein to DNA and shear DNA strands by sonication. Then use antibodies to immunoprecipitate target protein-DNA complex. Finally, reverse cross-link and label DNA. We can find motif by sequencing the labeled DNA.

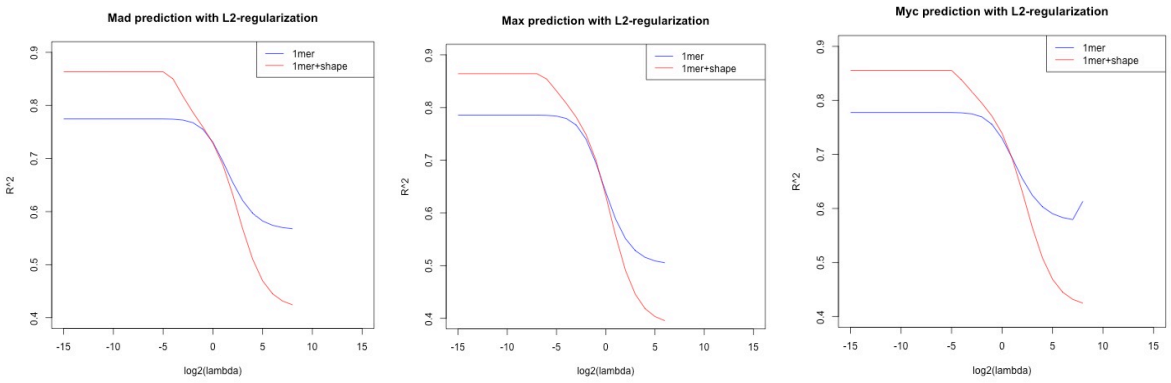
(c) The advantage of PBM and SELEX-seq is that they provide quantitative result, but because they are in vitro experiments, we cannot accurately assess the real interaction in living cell. The advantage of ChIP-seq is that this experiment is in vivo, so that it provides real binding events in living cell. However, ChIP-seq can only offer “yes” or “no” result, instead of quantifying the binding affinity.

(3) Vide the source code.

(4)

(a) We generate two feature vectors for each dataset using command “encodeSeqShape”. The two feature vectors encode “1-mer” and “1-mer+1-shape” respectively.

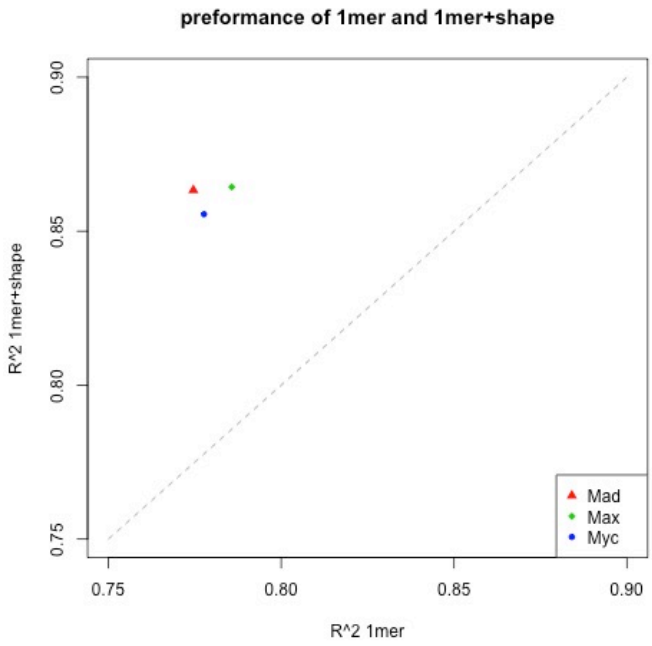
(b) We plot the average R^2 values for each dataset. Note that when lambda gets too large, the model cannot be trained. It is possible that all the parameters are shrank to 0.



The R^2 values of prediction models are listed as follow.

	R2.1mer	R2.1mershape
Mad	0.7745896	0.8633170
Max	0.7856984	0.8643521
Myc	0.7776688	0.8555111

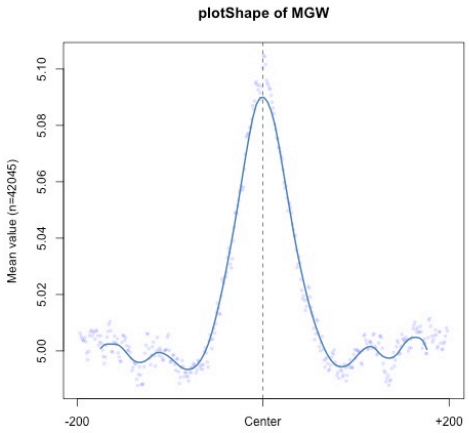
(5)
(a)

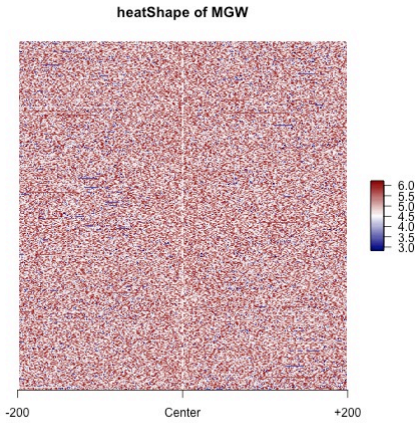


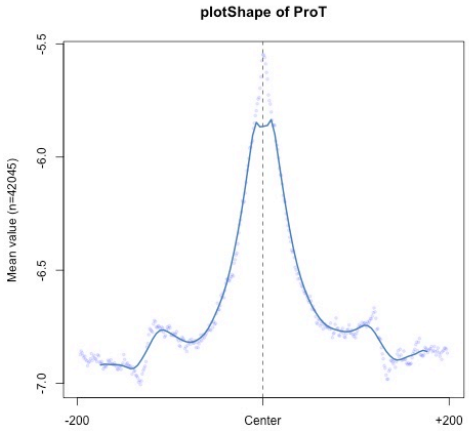
(b) We use Wilcoxon rank-sum test, which is a non-parametric statistical test for the difference of a particular measurement in two populations. Here, we don't have any assumption for the distribution of R^2 values, so the best choice is to use non-parametric statistical test without any additional assumption. The null hypothesis here is that mean of R^2 values of "1-mer" models are no less than the mean of R^2 values of "1-mer+1-shape" models.
The p-value we got is 0.05, which means rejection of the null hypothesis. R^2 values of "1-mer" models are generally less than R^2 values of "1-mer+1-shape" models.
(c) From the figure and the p-value, we know that the R^2 value of "1-mer+1-shape" is generally larger than that of "1-mer". R^2 reflects the amount of outcome can be interpreted by the input data. From the result, we can see that data with the predicted shape explains the experimental data better. The predicted shape does encode some important information.

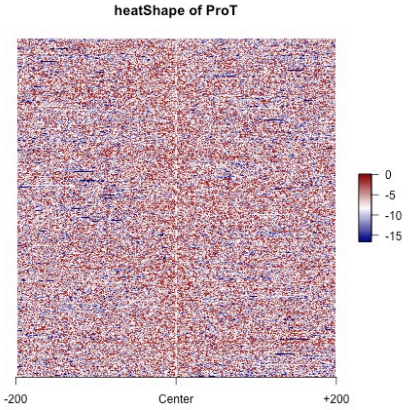
(6) Vide the source code.

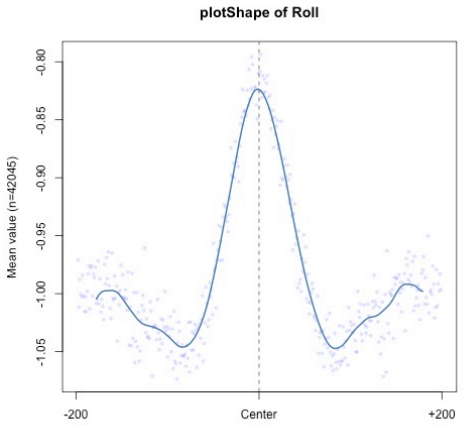
(7)
(a)

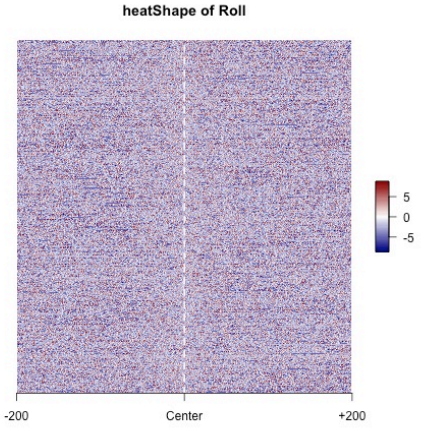

plotShape of MGW
Mean value (n=42045)
5.00
5.02
5.04
5.06
5.08
5.10
-200
Center
+200

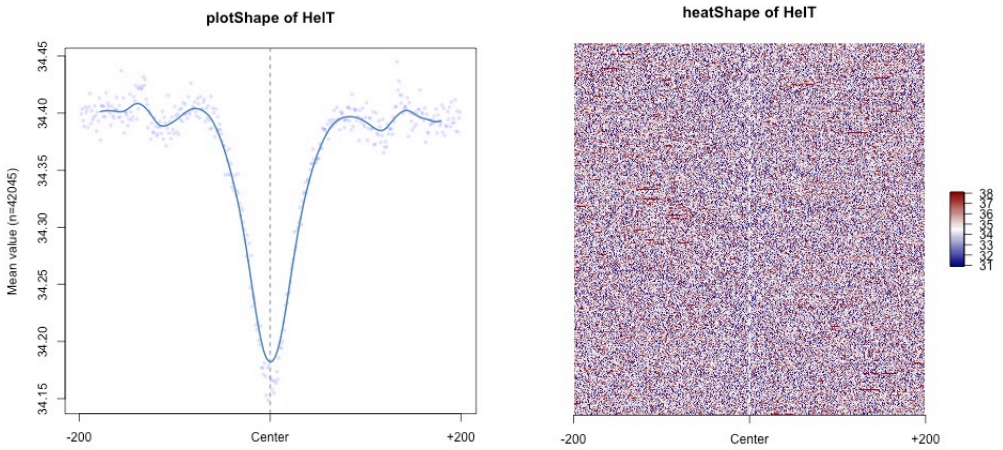

heatShape of MGW
6.0
5.5
5.0
4.5
4.0
3.5
3.0
-200
Center
+200


plotShape of ProT
Mean value (n=42045)
-7.0
-6.5
-6.0
-5.5
-200
Center
+200


heatShape of ProT
0
-5
-10
-15
-200
Center
+200

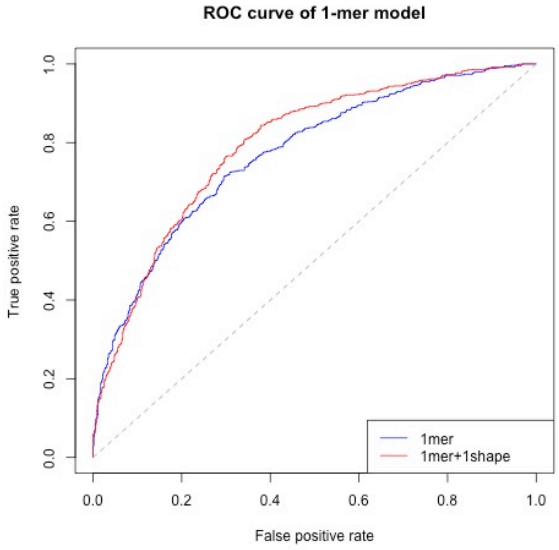

plotShape of Roll
Mean value (n=42045)
-1.05
-1.00
-0.95
-0.90
-0.85
-0.80
-200
Center
+200


heatShape of Roll
5
0
-5
-200
Center
+200



(b) The plots show the average shape properties of CTCF binding sites. We can see around the center of CTCF binding sites peaks are encountered in every plot. The minor groove width becomes larger at the center of binding sites, which might reflect the minor groove interaction between DNA and CTCF at the binding site. The propeller twist, roll, and helix twist of the binding site show the protein twists the DNA while binding it.

(8)
(a) Vide the source code.
(b)



AUC of the two models are listed as follow.
> auc1mer
[1] 0.7707413
> auc1mershape
[1] 0.7928581
(c) From the AUC values and ROC curves, we can see that the classifier considering shape information performs almost uniformly better than the classifier just taking sequence

information into account. Shape does encode some important information of the interaction between DNA and CTCF. More generally, we can suggest that DNA shape affect the binding of proteins.