MATH 577a-HW2

Amal Thomas

Q1.

Created git repository & made TsuPei collaborator.

Q2.

SELEX-seq is a technique in molecular biology to calculate relative affinities for a wide range of high, medium, and low-affinity binding sites. This technique combines classical protein-DNA SELEX (Systematic Evolution of Ligands by EXponential Enrichment) assays with massively parallel sequencing.

Advantages:

- higher sequencing coverage gives better inference
- cell-specific aptamers can be obtained without any knowledge about cell surface molecules on the target cells

Disadvantages:

- sequencing errors.

Protein-binding microarray (PBM) technology provides a rapid, high-throughput means of characterizing the in vitro DNA-binding specificities of transcription factors (TFs). In PBM high-density microarrays containing all 10-mer sequence variants is used to obtain a comprehensive binding-site measurements for any TF, regardless of its structural class or species of origin.

Advantages:

- Better quantitative information: the signal within each spot on the microarray corresponds to numerous DNA-protein binding events.

- Non-binding sequences can be identified.

Disadvantages:

- Repetitive PCR of bounded probes can lead to bias.

ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. In this technique we pull out those protein-DNA complex in which protein is bound to the target DNA, then the protein is degraded and purified DNA is sequenced.

Advantages:

- Increase in resolution and reduction in noise over ChIP-chip
- Genome coverage is not limited by the probe sequences fixed on the array.

Disadvantages:

- ChIP experiments cannot discriminate between different TF isoforms.
- Difficulty in creating large scale assays because specific antibodies has to be designed.

Q3.

Installed R-3.3.0, bioconductor, DNAshapeR & Caret

Cloned TsuPeiChiu github directory for the data.

Q4.

A) Feature vector created by the steps mentioned in script: Q4-5.R

B)

Table 1. showing average R^2 values for sequence model and sequence+shape model.

R^2 values		
	1-mer	**1-mer+Shape**
Mad	0.775	0.863
Myc	0.778	0.855
Max	0.785	0.864

It can be seen that the sequence + shape model performs clearly better than the 1-mer model when comparing the average R^2 values (Table 1).

Q5.

A)

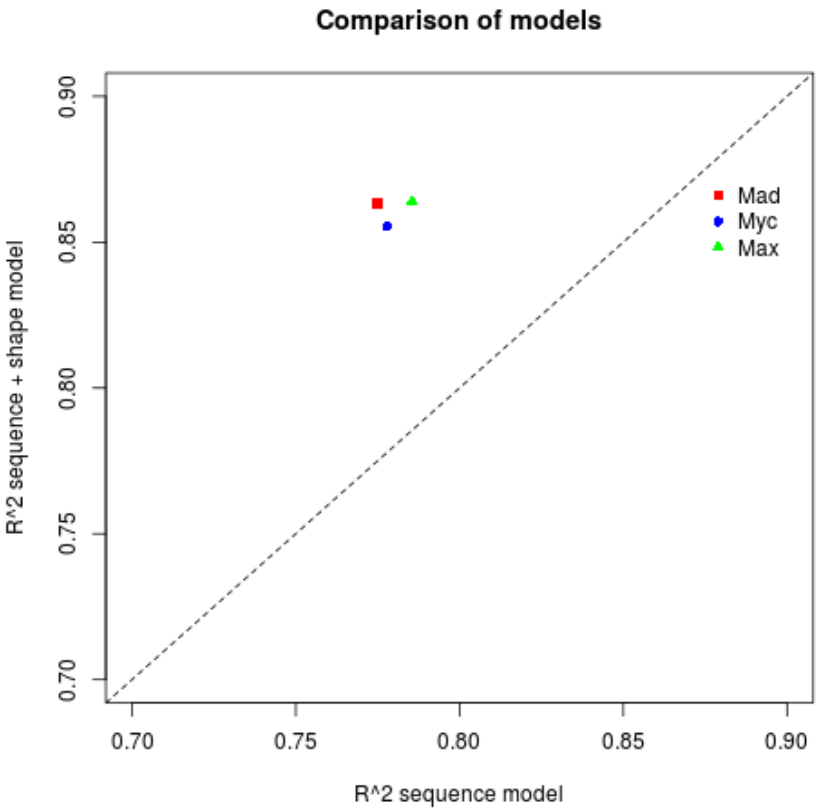

Figure 1: Graph comparing the R^2 values between two models.

B)

Mann-Whitney U test (also known as Wilcoxon rank-sum test) is used to find out the p-value (script Q4-5.R).

Command Used:

wilcox.test(y,x) # independent 2-group Mann-Whitney U Test

Null hypothesis: R^2 values under sequence model and sequence + structure model follows same distribution.

By performing the test, the resulting p-value obtained is 0.1. At 0.05 significance level, we conclude that the Null hypothesis is true that is there is no significant out performance by shape augmented model.

Note: But here we have to consider that the sample size is very small. This might affect our statistical analysis.

Q6.

Installed AnnotationHub and downloaded FASTA file using script Q6-7.RQ7.

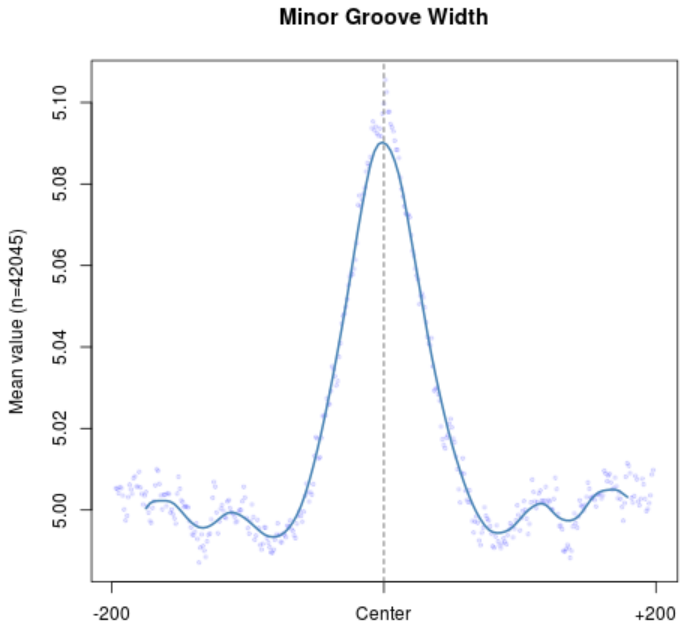

Figure 2. Plot showing Minor Groove width.

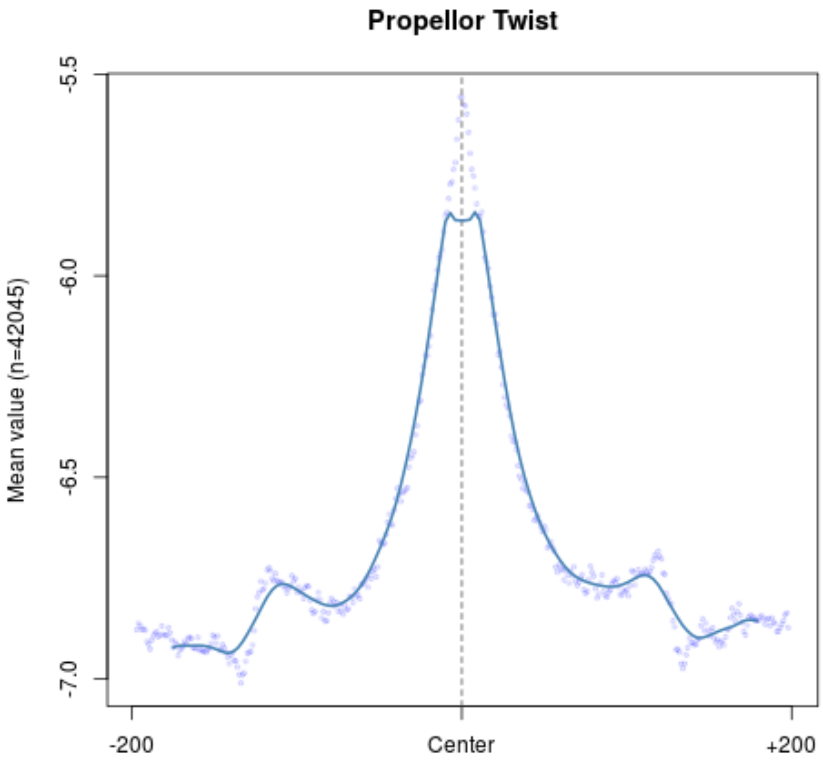

Figure 3. Propellor twist plot

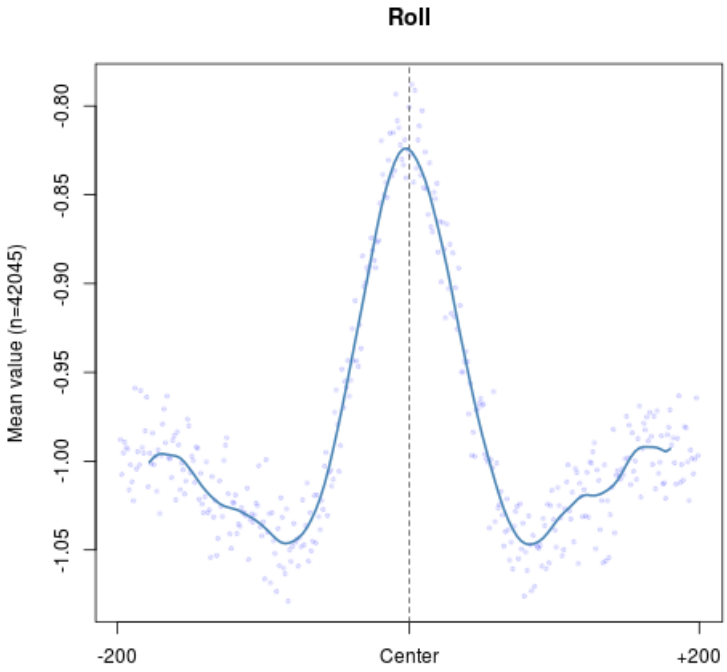

Figure 4: plot showing Roll

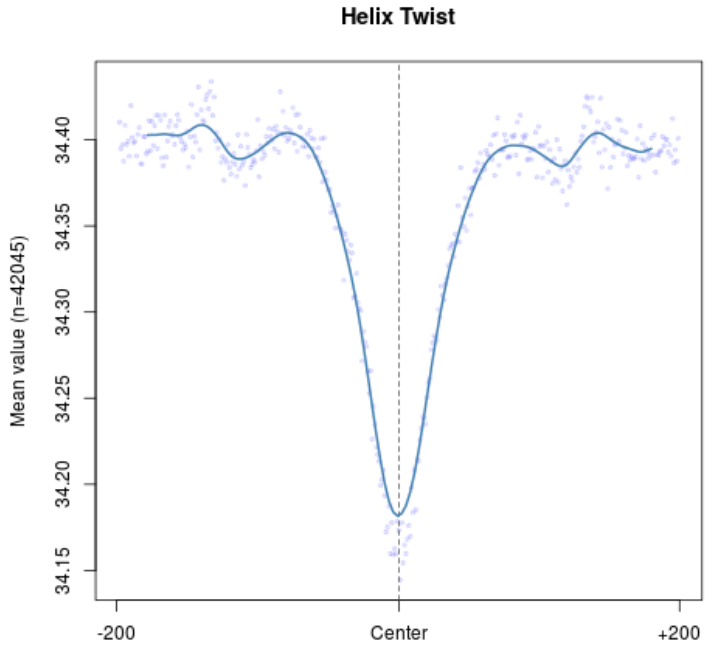

Figure 5: plot showing Helix twist.

B) Here we can see that as move from the center of binding each of the parameter values changes significantly. The value for minor groove width, propellor twist and roll is maximum at the center. This gives us the information that the minor groove in the binding site of DNA tends to be more open and thus become accessible to the TF/proteins to bind.

Q8.

A) Generate random genomic sequences: I resorted to BEDTools and Bash commands to do this. First genome size (genome.bed) and ctcf_cordinates bed files (ctcf_cordinates.bed) are generated as mentioned in the Q8.R. Then

1. Sort genome.bed using:

sort -k 1,1 -k2,2n genome.bed > sorted_genome.bed

2. Sort ctcf_cordinates.bed using:

sortBed -i ctcf_cordinates.bed > sorted_ctcf_cordinates.bed

3. Find complement of ctcf_cordinates by:

bedtools complement -i sorted_ctcf_cordinates.bed -g sorted_genome.bed > complement_cordinates.bed

4. Get random 1000 coordinates with size 30 using:

bedtools random -g complement_cordinates.bed -n 1000 -l 30 > random_cordinates.bed

5. To get fasta file for the random-cordinates. Downloaded mm10 file from UCSC and converted twoBit file to fasta using:

twoBitToFa mm10.2bit mm10.fa

6. Extracted the fasta sequences using getFastaFromBed

bedtools getfasta -fi mm10.fa -bed random_cordinates.bed -fo unbound.fa

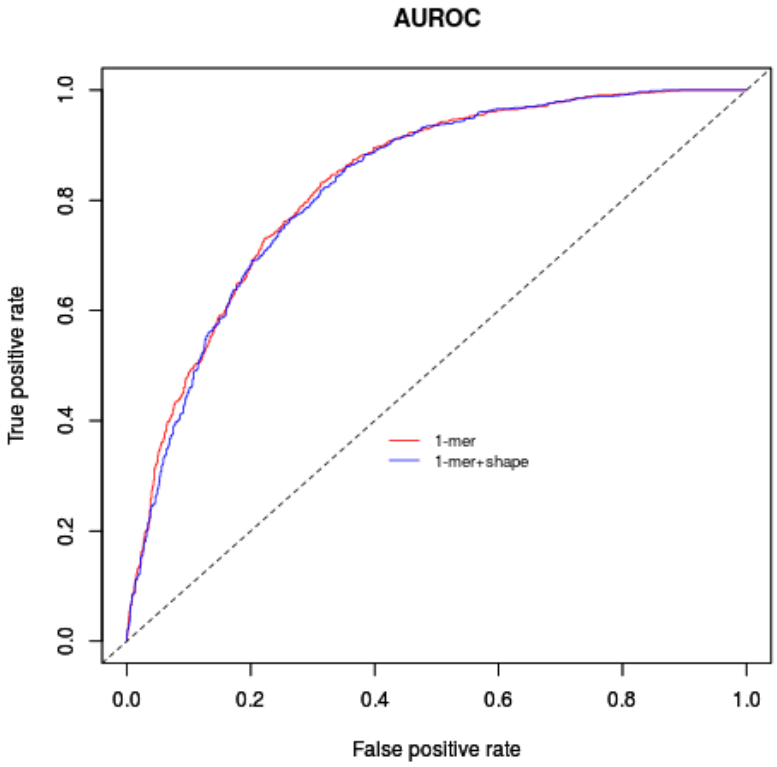

Figure 6: AUROC curve

B)

AUROC plot shown in figure 6.

Code: Q8.R

Logistic regression model was built for 1mer and 1mer+shape.

- Area under the curve for sequence model: 0.83049
- Area under the curve for sequence + shape model: 0.825897

C)

The area under the curve for both the model are almost same. When we include the shape factor, we expect to get a better prediction for the binding sites. But here the small sample size of 1000 sequences each of length 30 nt might have affected our prediction.