Hao Zheng
BISC 481
Assignment #3

1.
See GitHub

2.

(a)

SELEX-seq: Systematic evolution of ligands by exponential enrichment
In a random pool of oligonucleotides, immobilized target proteins are used for specific specific selection. The bounded oligonucleotides sequences are amplified by PCR and undergo selection again.

PBM: Protein binding microarray
With a dsDNA microarray, epitope-tagged transcription factors are used to bind dsDNA, which are then labelled with fluorophore-tagged antibodies. The microarray is scanned and "sequence vs transcription factor binding intensity" is generated.

(b)

ChIP: Chromatin immuneprecipitation
DNA is crosslinked and sonicated. Specific antibodies are added and then DNA is immunoprecipitated. Crosslinks are reversed while the DNA is labelled and then microarray sequence is generated. Specific motifs are ultimately discovered.

(c)

	PBM	SELEX-seq	ChIP
Advantage	can be utilized for quantitative measurement and massively parallel testing of protein function	greater coverage of selected DNA, fewer rounds of selection, and the biophysical sequence-to-affinity model, captures more than just high affinity binding sites, and thus provides a more complete view of the binding preferences for a TF or TF-complex	higher resolution, less noise, and greater coverage
Disadvantage	relatively low resolution	selection of a aptamer than may not have any inhibitory activity towards the target	can only offer classification, cost and availability

3.
See codes

4.
(a) See codes

(b)

R^2	1-mer+shape	1-mer
Mad	0.8626035	0.7746128
Max	0.8648306	0.7856608
Myc	0.8549286	0.778698

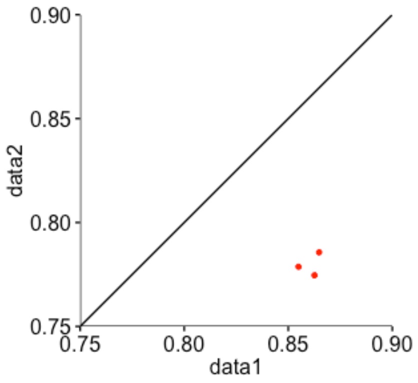
R^2 is a number that indicates how well data fit a statistical model.

5.

(a) Plot:

Data1 represents “1-mer” R^2

Data2 represents “1-mer+shape” R^2



For the same sequence, 1-mer+shape data offer higher R^2.

(b) Discussion:

Feature vectors can be generated by DNAshapeR for a user-defined model, which can consist of sequence feature, such as 1-mer, or shape features. In 1-mer features, sequence is encoded in binary numbers at each nucleotide position, while shape features are values for the MGW, Roll, ProT and HelT. Then a feature matrix is the result from feature encoding of multiple sequences. Ultimately a machine learning package, such as caret package, can be used to train a multiple linear regression model

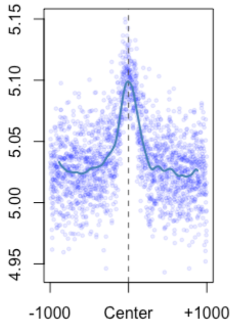
In this specific case, the R^2 values of “1-mer+shape” are generally larger than “1-mer” vector, indicating a higher accuracy in the combined feature vectors.

6.

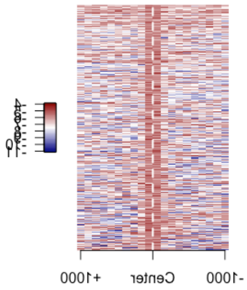
See codes

7. (a) Plots:

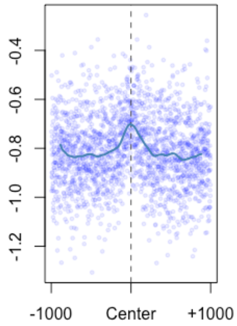
Minor groove width:



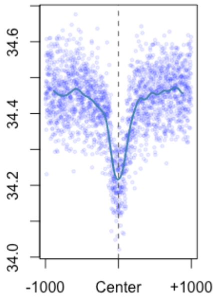
Propeller twist:



Roll:



Helix twist:



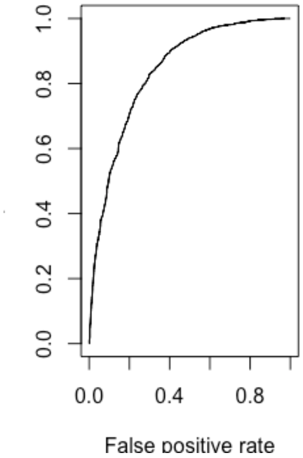
(b) Discussion:

The DNAshape method can define a vector of minor groove width, roll, propeller twist, and helix twist at each nucleotide position, while MGW and ProT define base-pair parameters whereas Roll and HelT represent base pair-step parameters. Depending on DNAshapeR, DNA shape features can be predicted from FASTA files. It can also be used to generate graphical representations and visualized in scatter plots.

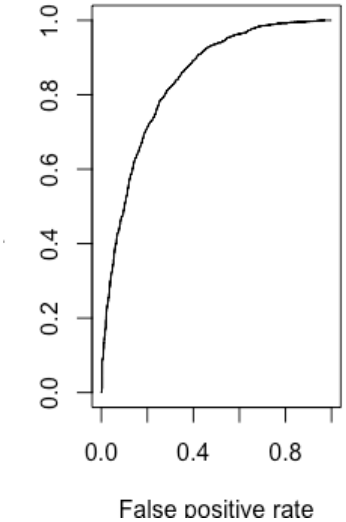
8.

(a) Plots of ROC curves for 1-mer and 1-mer+shape:

1-mer:
AUC score: 0.8395175



1-mer+shape:
AUC score: 0.8393532



The difference in R^2 values is not significant between “1-mer” and “1-mer+shape”, which means shape does not play an important role in this case.

(b) Discussion:

AUC represents “area under ROC curve”, which is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, while ROC means “receiver operating characteristic”. ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from the cost context or the class distribution, and it is related in a direct and natural way to cost/benefit analysis of diagnostic decision making