BISC481

Homework 3

10/18/2016

Anastasiia Sadybekov

Q2.

(a) **SELEX-seq** and **PBM** are methods to analyze protein-DNA binding in vitro.

SELEX-seq method combines classical protein-DNA SELEX (Systematic Evolution of Ligands by EXponential enrichment) assays with massively parallel sequencing. The library of potential binding sites, that are flanked by defined primer docking sites, is created. Investigated protein is added to the DNA library. DNA bound by the complex is then separated from unbound DNA and the bound DNA is then amplified by PCR and used for subsequent rounds of DNA binding and selection.

PBM use microarrays with double-stranded DNA probes to measure the fluorescence of alphaGST-tagged proteins bound to their sequence-specific binding sites on the probes. PBM uses arrays of all possible ten-base-long sequences of the nucleotides. Investigated protein is added to the array, which is then washed to minimize non-specific binding. The remaining protein is quantified with a fluorescent antibody. Detecting of fluorescence group reveal oligonucleotides that bind protein.

(b)

ChIP-seq – method to analyze protein-DNA binding in vivo. DNA-binding protein is crosslinked to DNA in vivo. Chromatin is sheared into small fragments. Then an antibody specific to the protein of interest is used to immunoprecipitate the DNA-protein complex. The crosslinks are reversed and the released DNA is assayed to determine the sequences bound by the protein.

(c)

	Advantages	Disadvantages
Selex-seq	No limit to the size of the binding site Can be used for large protein complexes Quantity of binding data is limited only by the depth of sequencing	Relatively expensive The initial pool is biased, further PCR amplification only brings additional biases Provides integer-valued, poisson-distributed sequence read counts
PBM	Fast Relatively inexpensive Provides real-valued measurements of binding	Difficult to model large binding sites (>10 base pairs)
ChIP-seq	Not limited limited by array design High spatial resolution	Cost The quality can be low

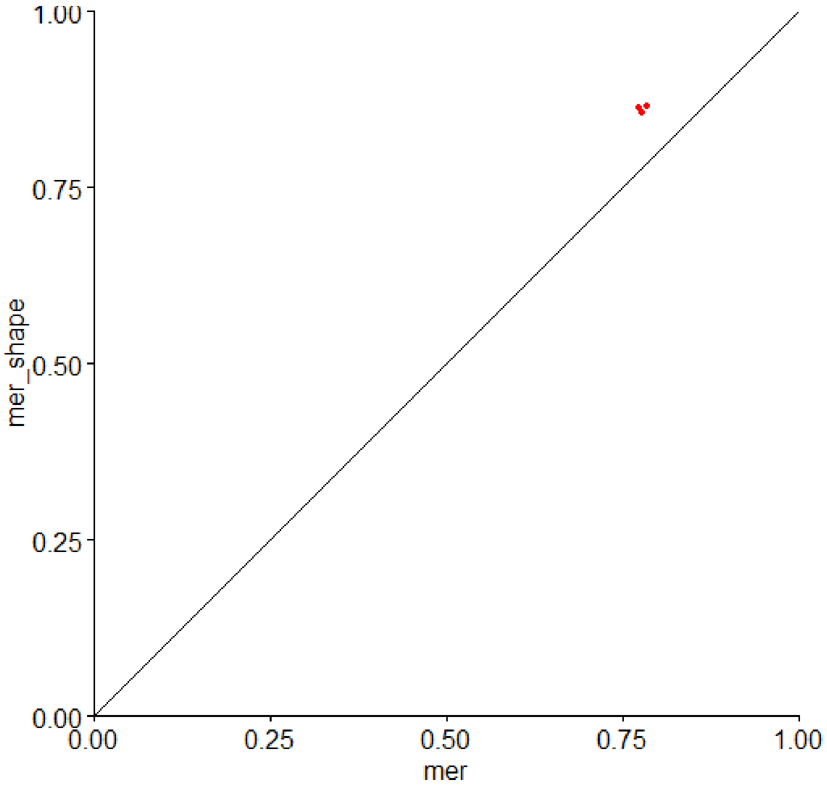
In vivo method provides only qualitative data: discriminates between binding and non-binding. While in vitro methods give quantitative data: the affinity can be found.

Q4.

R^2

	Mad	Max	Myc
1-mer	0.775	0.785	0.778
1-mer+shape	0.863	0.865	0.855

Q5.

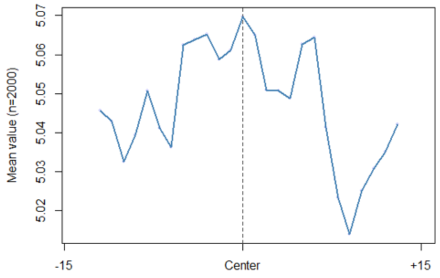


Performance comparison for Mad, Max and Myc datasets.

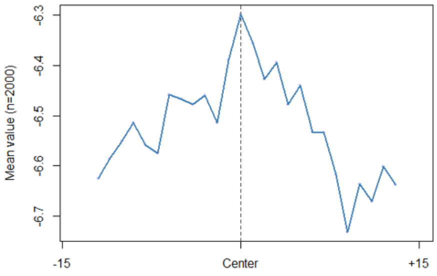
Discussion: From the obtained data we can see that DNA shape contributes to the DNA binding specificity of all three datasets. Incorporation of DNA shape features into the binding specificity model allowed improvement of R^2 for all three datasets: Mad, Max and Myc.

Q7. Bound

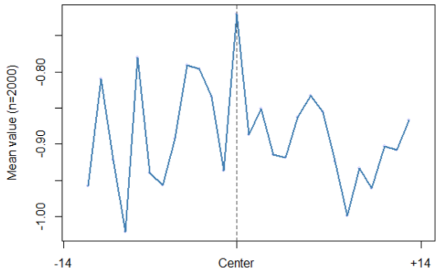
MGW



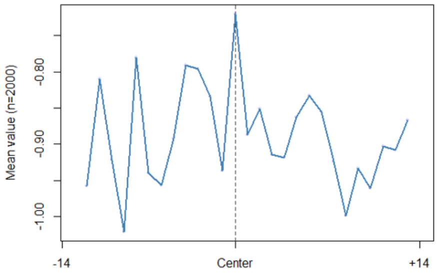
ProtT



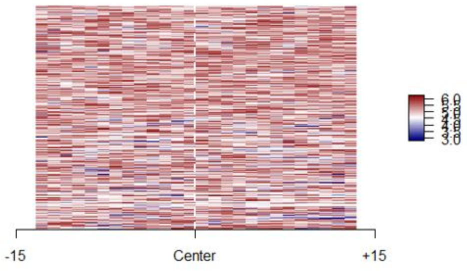
Roll



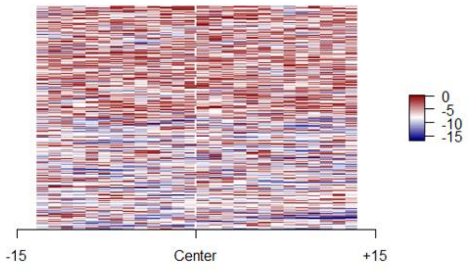
HelT



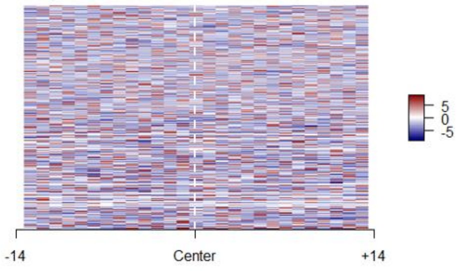
MGW



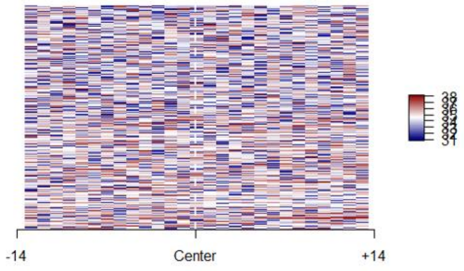
ProtT



Roll

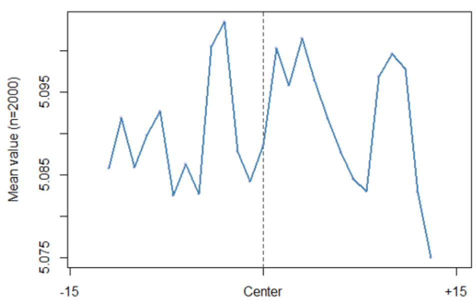


HelT

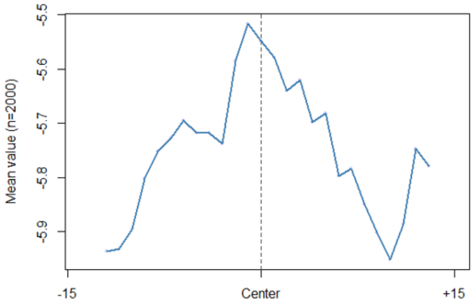


Bound_30

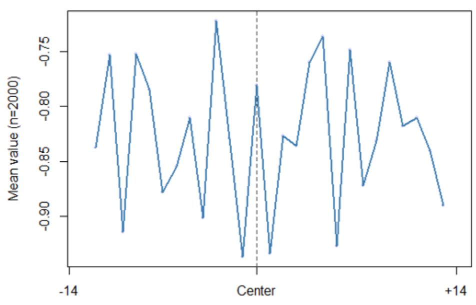
MGW



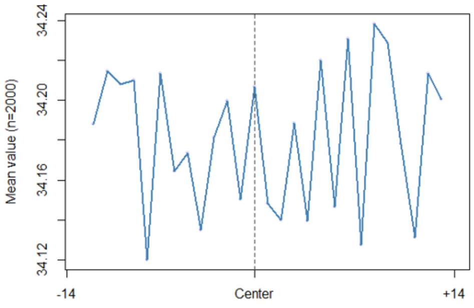
ProT



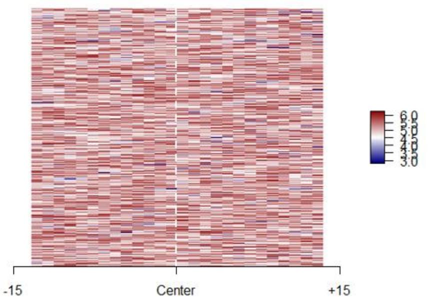
Roll



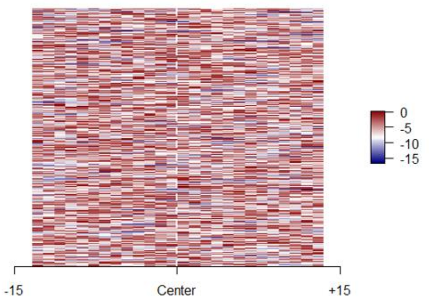
HelT



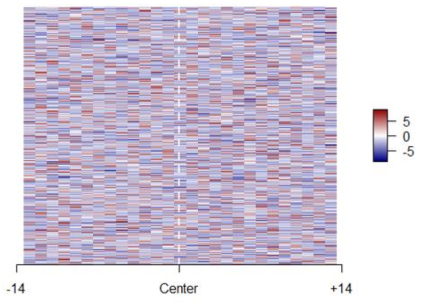
MGW



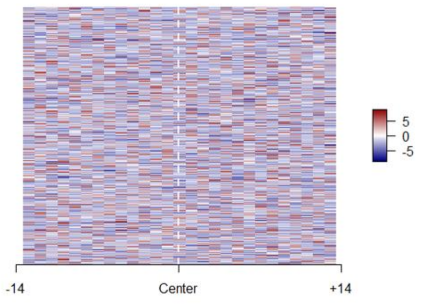
ProT



Roll

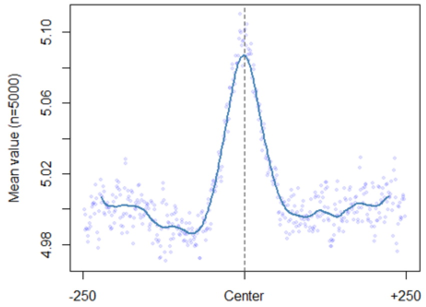


HelT

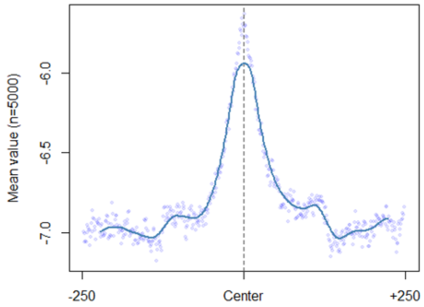


Bound_500

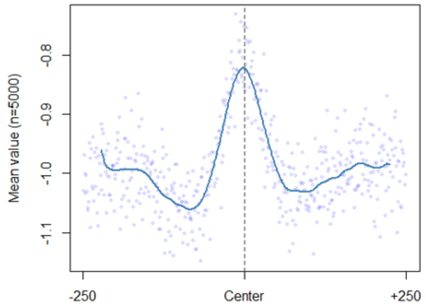
MGW



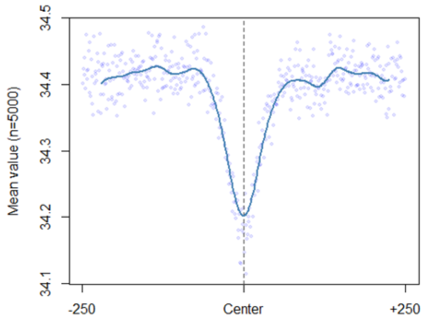
ProT



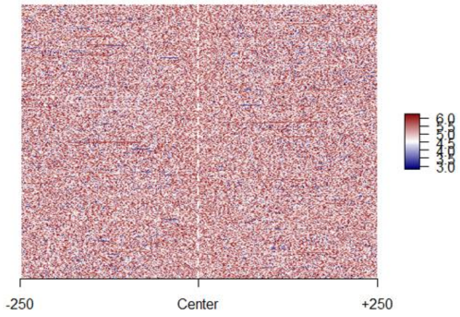
Roll



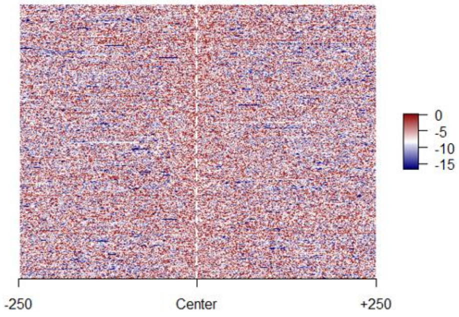
HelT



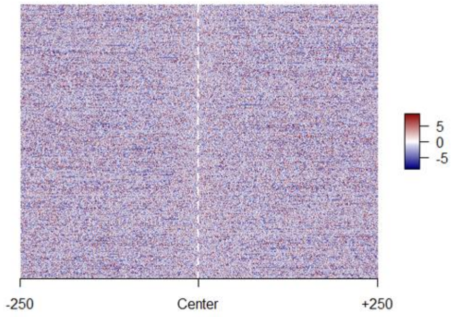
MGW



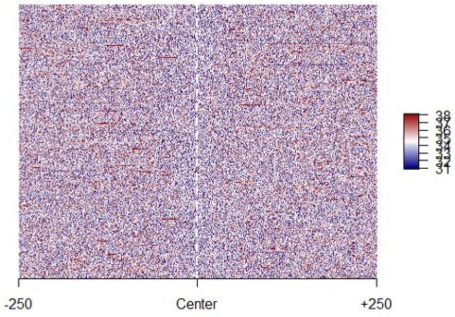
ProT



Roll

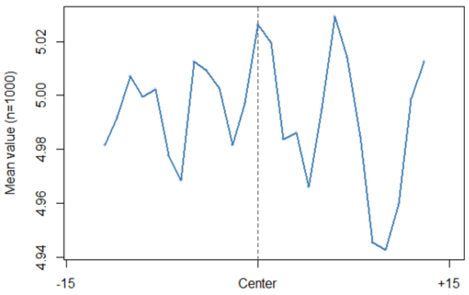


HelT

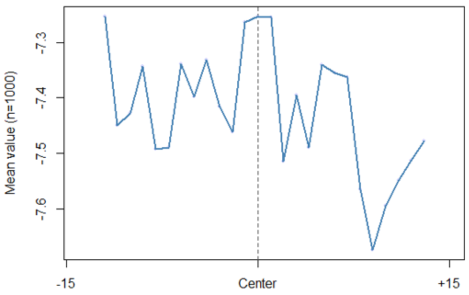


Unbound

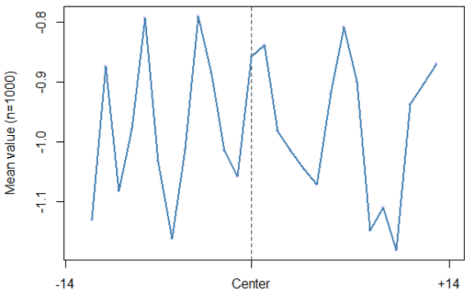
MGW



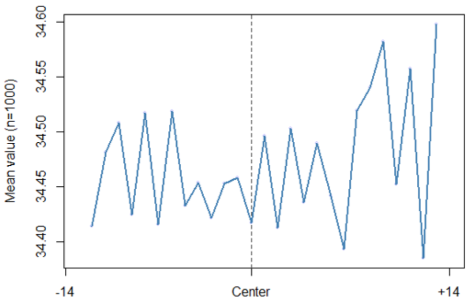
ProT



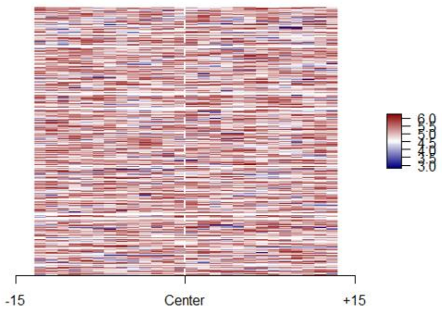
Roll



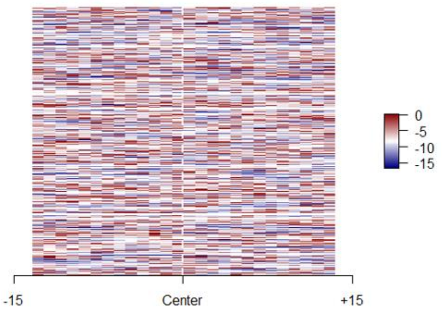
HelT



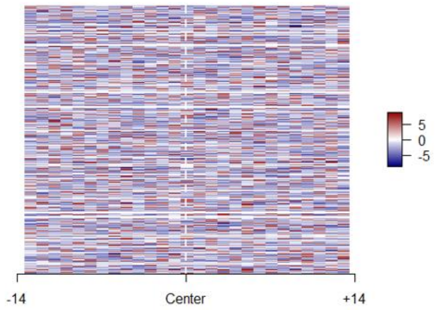
MGW



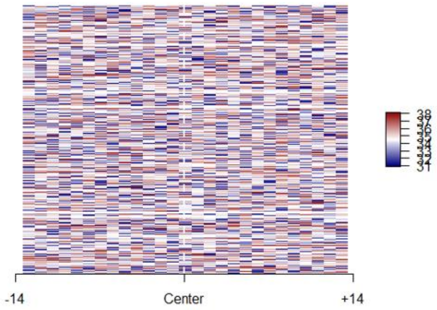
ProT



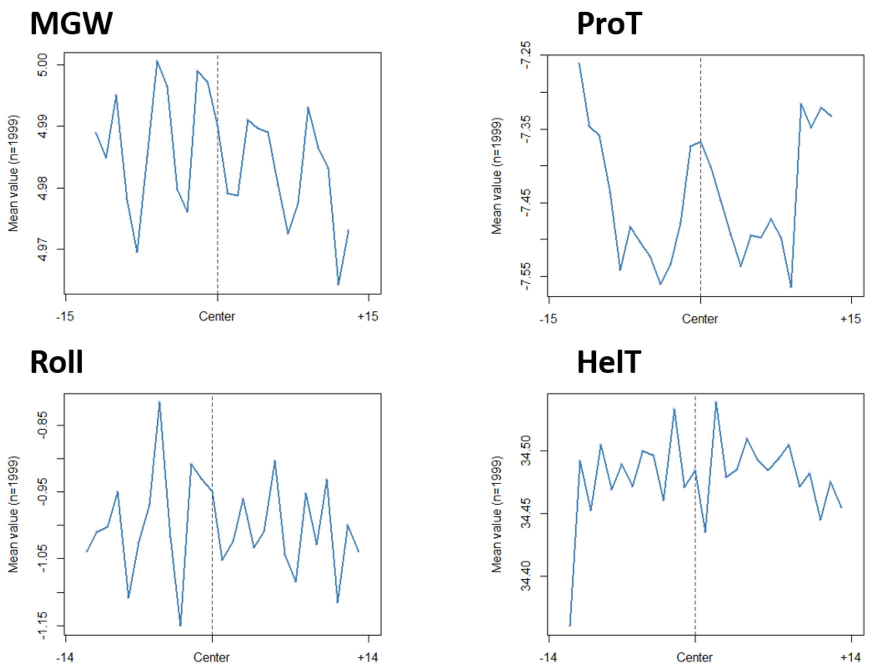
Roll

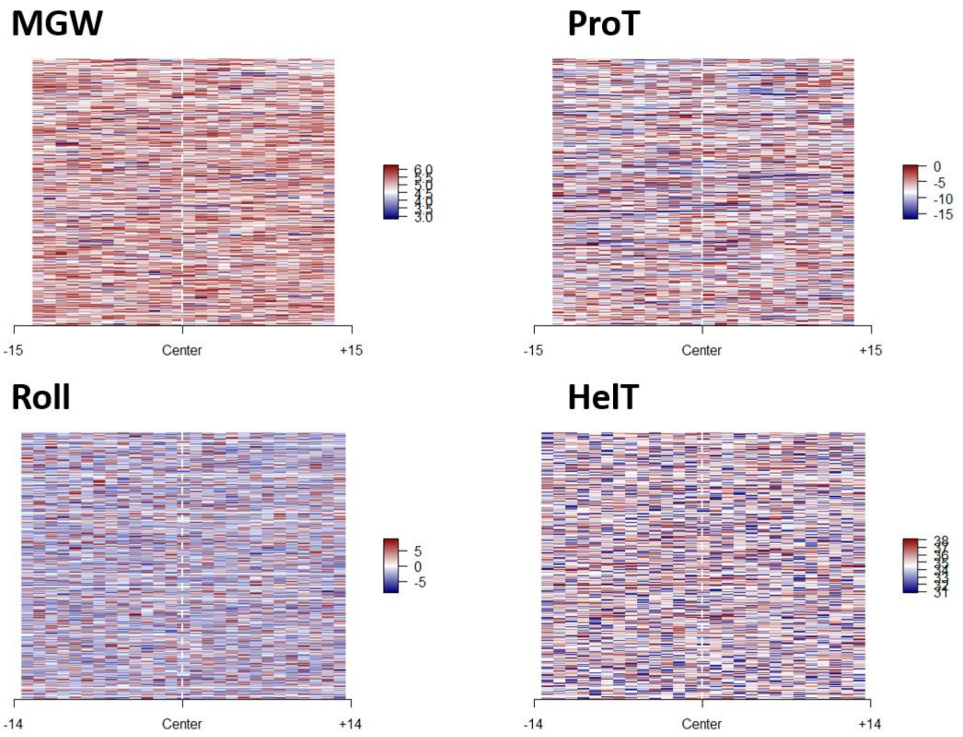


HelT



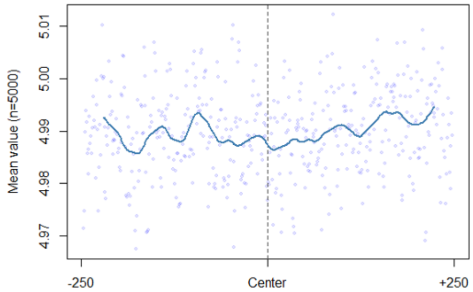
Unbound_30


MGW
Mean value (n=1999)
5.00
4.99
4.98
4.97
-15
Center
+15
ProT
Mean value (n=1999)
-7.25
-7.35
-7.45
-7.55
-15
Center
+15
Roll
Mean value (n=1999)
-0.85
-0.95
-1.05
-1.15
-14
Center
+14
HelT
Mean value (n=1999)
34.50
34.45
34.40
-14
Center
+14

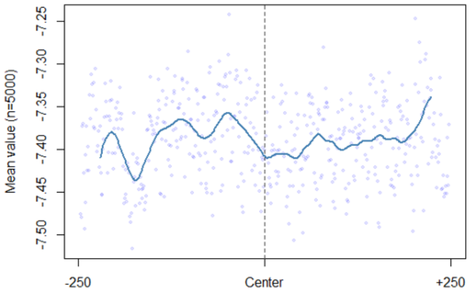

MGW
6.0
3.0
-15
Center
+15
ProT
0
-5
-10
-15
-15
Center
+15
Roll
5
0
-5
-14
Center
+14
HelT
38
31
-14
Center
+14

Unbound_500

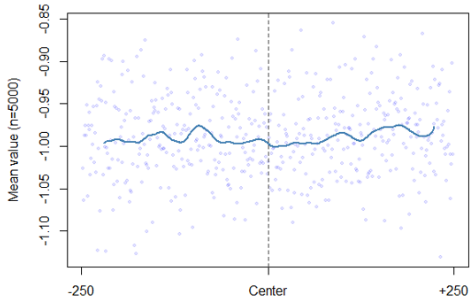
MGW



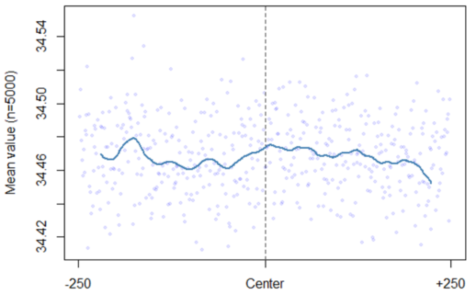
ProT



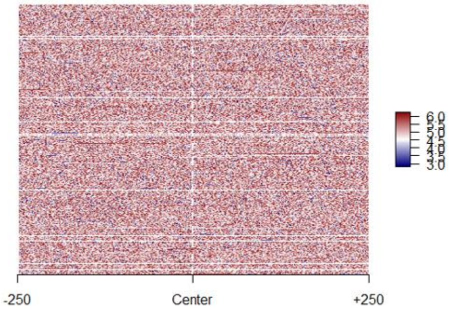
Roll



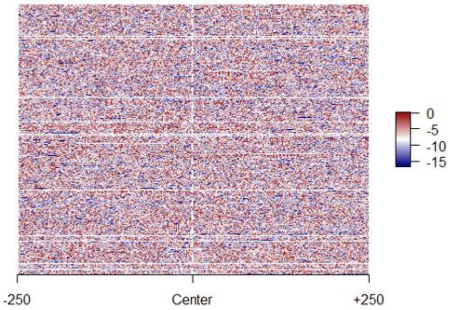
HelT



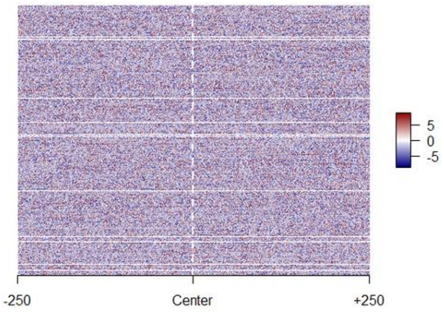
MGW



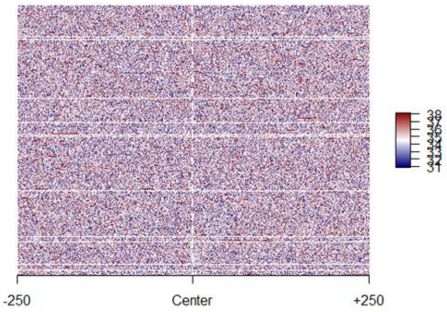
ProT



Roll



HelT

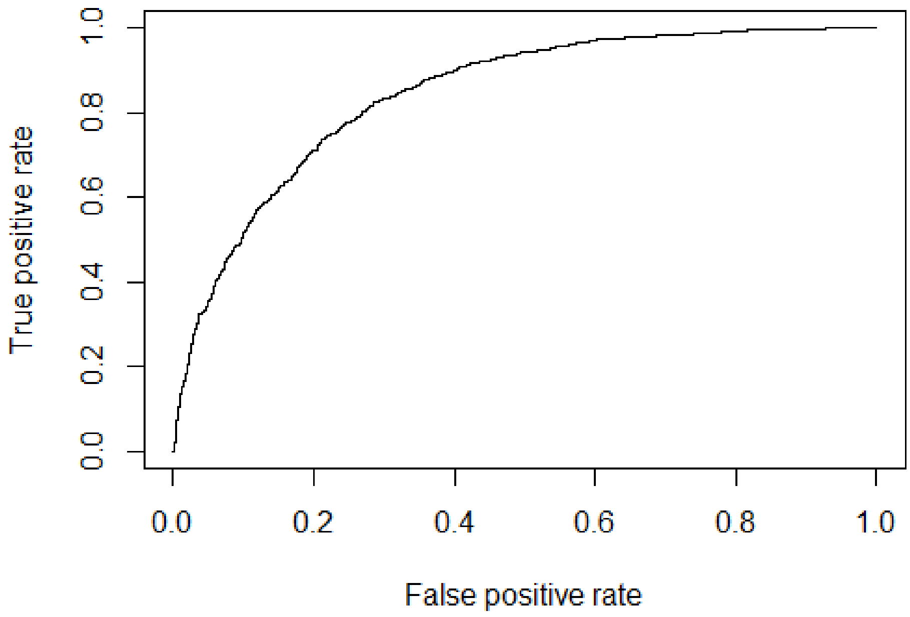


Discussion:

- Comparison of bound_500 and unbound_500 data show that binding site of the DNA has specific shape parameters distinct from the average parameters for unbound DNA: (1) Increased minor groove width; (2) Increased value of propeller twist parameter; (3) Increased value of Roll parameter; (4) Decreased value of helix twist parameter. These results suggests that DNA shape plays important role in binding specificity.
- Comparison of bound_30 and bound_500 or unbound_30 and unbound_500 shows that there is not enough data in bound_30 and unbound_30 to characterize shape features that play role in binding specificity. In the same time bound_500 and unbound_500 data provide enough information for shape characterization.

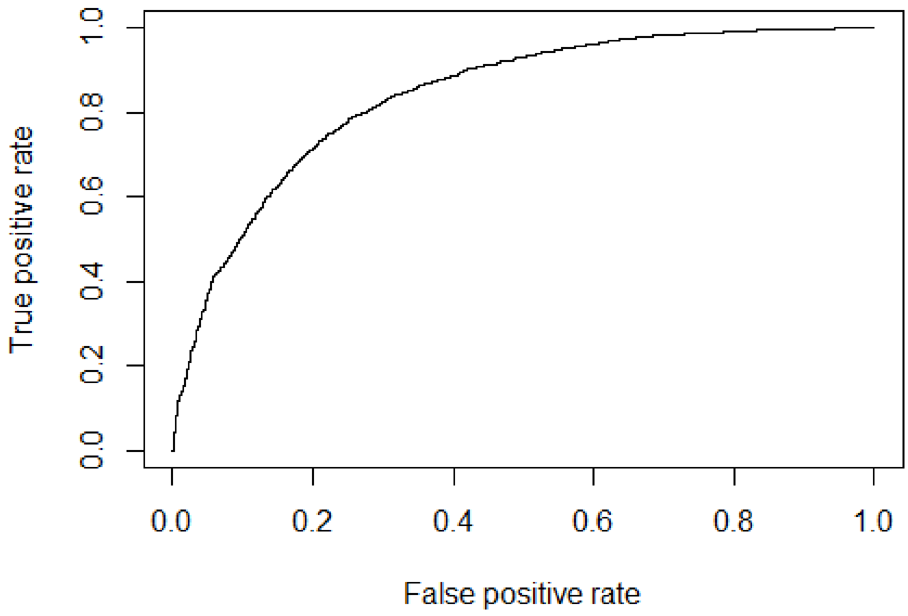
Q8.

Bound_30/Unbound_30: 1mer



AUC (1-mer) = 0.843

Bound_30/Unbound_30: 1mer+1shape



AUC (1-mer + 1-shape) = 0.840

Discussion:

- Both models are better than random, as both of them have AUC score higher 50.
- Although use of sequence + shape features was expected to provide better models, comparison of logistic regression models for sequence and sequence + shape features shows that models based on sequence features only are better than models based on sequence and shape features. This fact can be explained by the overfitting, where the sequence + shape model describes random error or noise instead of the underlying relationship.