BISC 577A HW2

Linda Ding

Q1

https://github.com/luna-stargazer/BISC577

Q2

Description of methods

SELEX

The general strategy is to create a library of potential binding sites, which may be from randomly synthesized DNA or created from genomic sequences. Purified TF is added to the library of DNA sites and the bound and unbound sequences are separated by various means. Typically, after several rounds, the selected sites would be cloned and sequenced, often obtaining fewer than 100 independent sites.

PBM

uses arrays of over 44,000 spots that contain all possible ten-base-long DNA binding sites once on each array, which means that every eight-base-long sequence occurs 32 times, taking both orientations into account. A TF, either purified from cells or synthesized *in vitro*, is added to the array, which is then washed to remove nonspecific binding. The amount of protein binding to any specific DNA spot is determined with a fluorescent antibody to the protein.

ChIP-Seq

Genomic sequences containing the mark of interest are enriched by binding soluble DNA chromatin extracts (complexes of DNA and protein) to an antibody that recognizes the mark. An analysis of the precipitated DNA through sequencing of the precipitated DNA then follows.

Comparing different methods:

SELEX

Pros: capable of determining important aspects of the binding specificity, including the consensus sequence and the relative variability in affinity for different bases at different positions within the binding sites. Also no inherent limit to the length of the binding site that can be selected.

Cons: if multiple rounds of selection are used it is not straightforward to determine binding energy distri- butions directly.

PBM

Pros: this technology has made possible large-scale, high-throughput analyses to collect information that previously was much more laborious to acquire

Cons: the optimal method for modelling the specificity of TFs based on PbM data remains an open question, with different methods some- times leading to different conclusions (see the section on computational modelling).

ChIP-Seq

Pros:can provide a framework for the regulatory network, indicating which factors are likely to regulate which genes.

Cons: the resolution of the binding locations is not sufficient to identify the binding sites precisely, only to indicate a region of 100 or more base pairs in which it resides

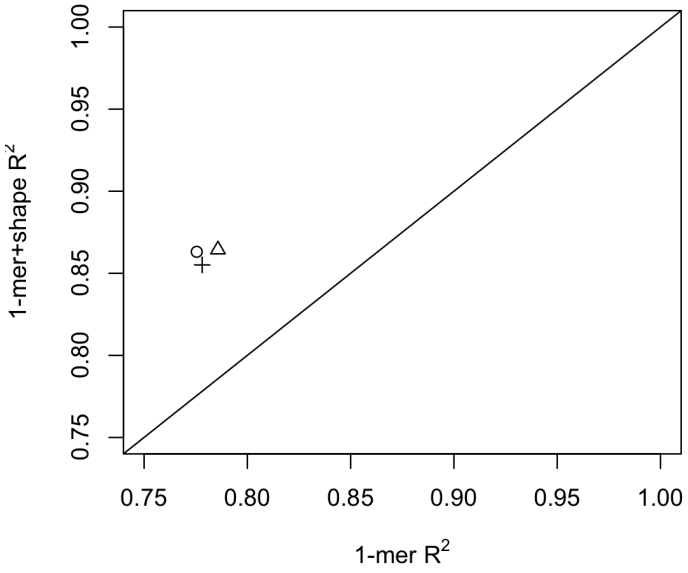
Q3.

Downloaded and installed

Q4.

Data	1-mer R^2	1-mer+shape R^2
Mad	0.775	0.863
Max	0.786	0.864
Myc	0.778	0.855

Q5.

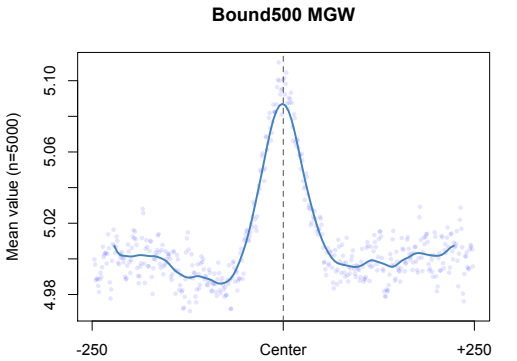


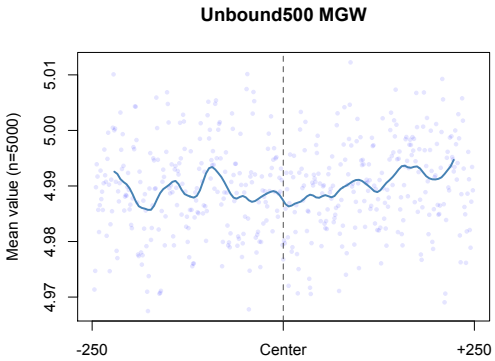
The shape-augmented (1mer+shape) model outperforms the sequence-only (1mer) model; the 1mer+shape model leads to consistent improvements

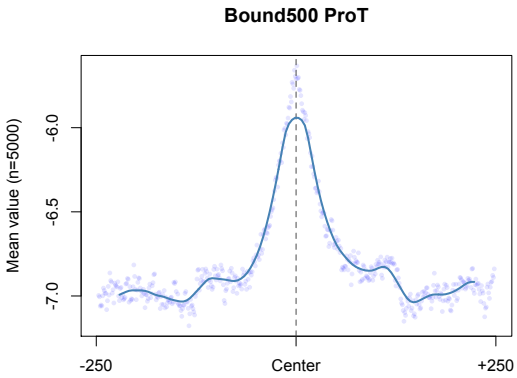
Q6.

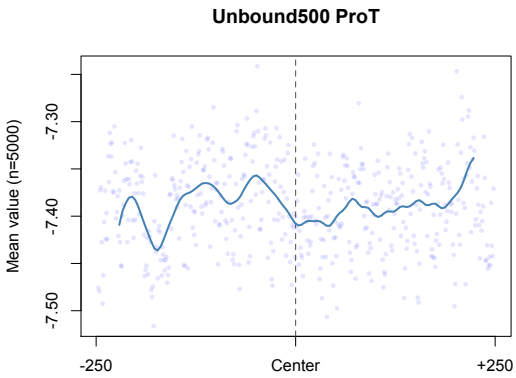
Files exist, packages installed

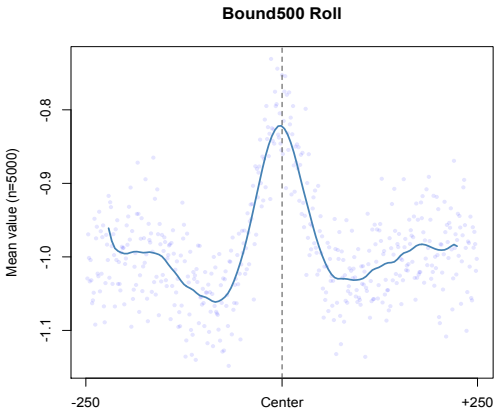
Q7.

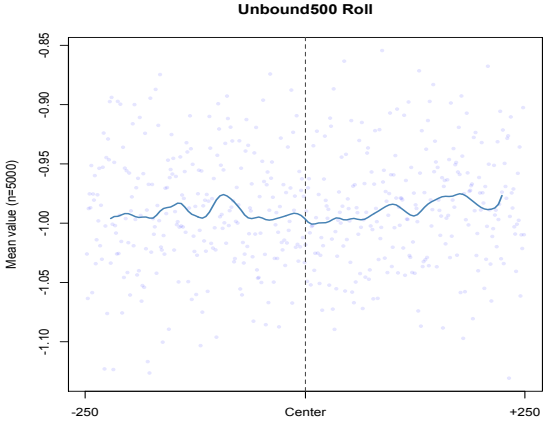


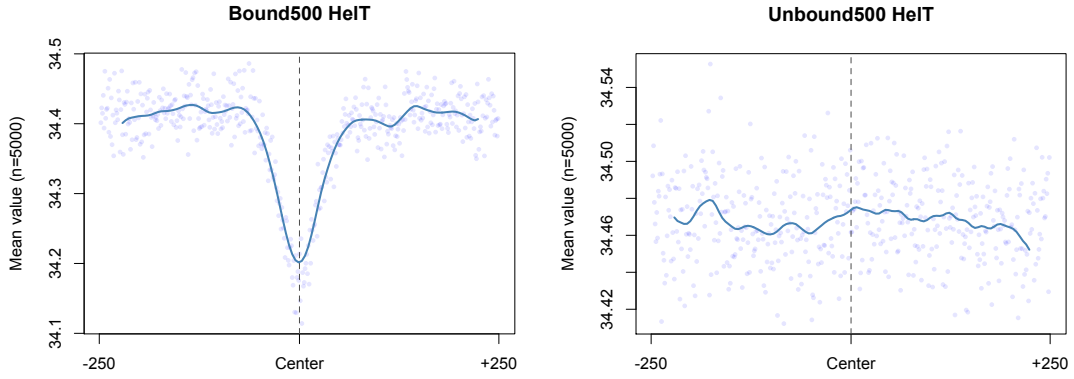






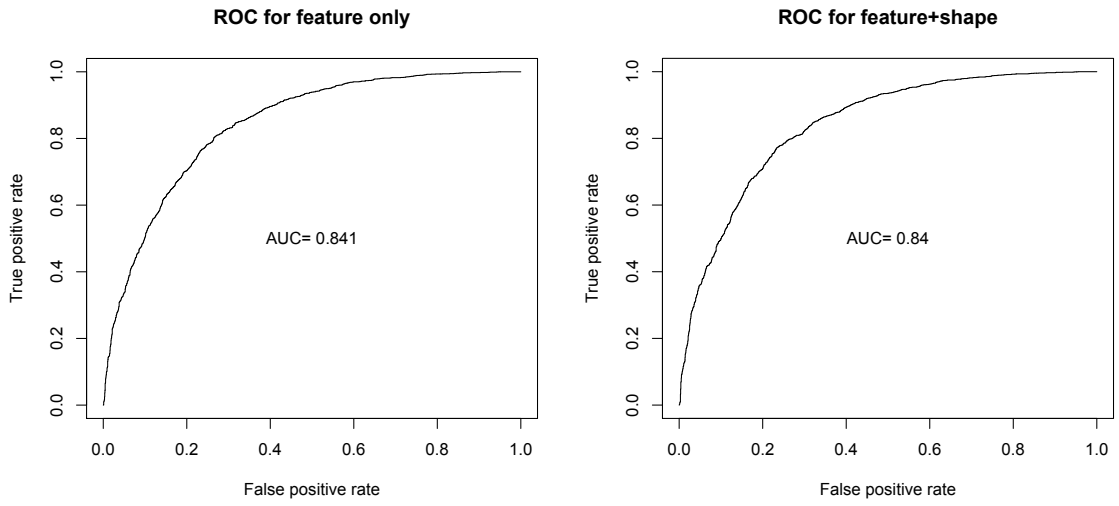






The protein prefers binding sites with larger than average minor groove width (MGW) , propeller twist (ProT), larger roll, and smaller than average helical twist

Q8.



Adding shape as a predictor does not appear to improve the model fit on the ChIP-Seq data. However AUC score is quite high for both models.