Assignment 2

Anna Nadtochiy

Q2. High-throughput binding essays

A in vitro: SELEX-seq and PBM

SELEX-seq and PBM can be used to determine the relative affinities to any DNA sequence for any transcription factor or multiprotein complex (target).

In SELEX-seq experiment, a large DNA oligonucleotide library is synthesised by randomly generating sequences of fixed length. Constant 5' and 3' ends are added to these sequences as primers for PCR. This library is sequenced and exposed to the target. DNA molecules, that do not bind the target are removed. The bound sequences are eluted, amplified by PCR and sequenced.[2]

This step is repeated multiple times and the stringency of the elution conditions is increased to identify the tightest-binding sequences.[3] The relative binding affinity for each K-mer is estimated based on the relative fold-enrichment with new rounds of selection.

In PBM experiment, target binds to a double-stranded DNA microarray. Similarly to SELEX-seq initial library, this micro array can be designed to contain all possible K-mers.

The target-bound array is labeled with a fluorophore-conjugated antibody specific to the target. Flourescent signal intensities from a target-bound microarray provide the relative amounts of target bound to each of the sequences and thus, provide the relative affinity of the target for all k-mers.

B in vivo: ChIP-seq

In ChIP-seq experiment, DNA and associated proteins on chromatin in living cells are crosslinked. The DNA-protein complexes are then sheared into short DNA fragments and DNA fragments associated with the proteins of interest are selectively immunoprecipitated.

Then, this DNA fragments are purified and sequenced. Enrichment of specific DNA sequences represents regions on the genome that the protein of interest is associated with in vivo.

C Comparison

Although, in vivo binding gives a picture of the processes, happening in the cell, it is affected by chromatin structure, nucleosome positioning and co-factors, thus

the accessible sites may not cover all possible k-mers. [1]

In vivo binding, depends on direct target–DNA interaction and allows sampling of the full spectrum of DNA k-mers. But it is unclear how accurate these models are in predicting in vivo binding. Also, these models may include technology specific biases.

Q4. Build prediction models for in vitro data

Dataset	R^2_{1-mer}	$R^2_{1-mer+shape}$
Mad	0.7753123	0.8628297
Max	0.7852302	0.8640685
Myc	0.7789938	0.8543180

Table 1: Coefficient of determination, R^2, for "1 mer" sequence-only model and "1 mer + shape" shape-augmented model with respect to the different data-sets.

Q5. High-throughput in vitro data analysis

There is a large improvement on R^2 for a shape-augmented model, "1 mer+shape", compared to the sequence-only model, "1 mer", on all three data sets that were tested (the human bHLH TFs Mad1 (Mad), Max, and c-Myc (Myc)), see Fig.1.

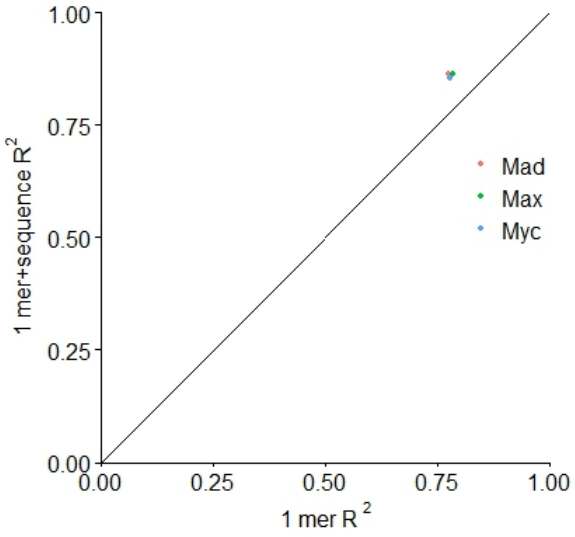

Figure 1: Coefficient of determination, R^2, for "1 mer" sequence-only model and "1 mer + shape" shape-augmented model for the data sets Mad, Max and Myc.

Q7. High-throughput in vivo data analysis

Bound regions have a specific DNA shape: wider minor groove, increased propeller twist and roll, decreased helix twist, see Fig.2.

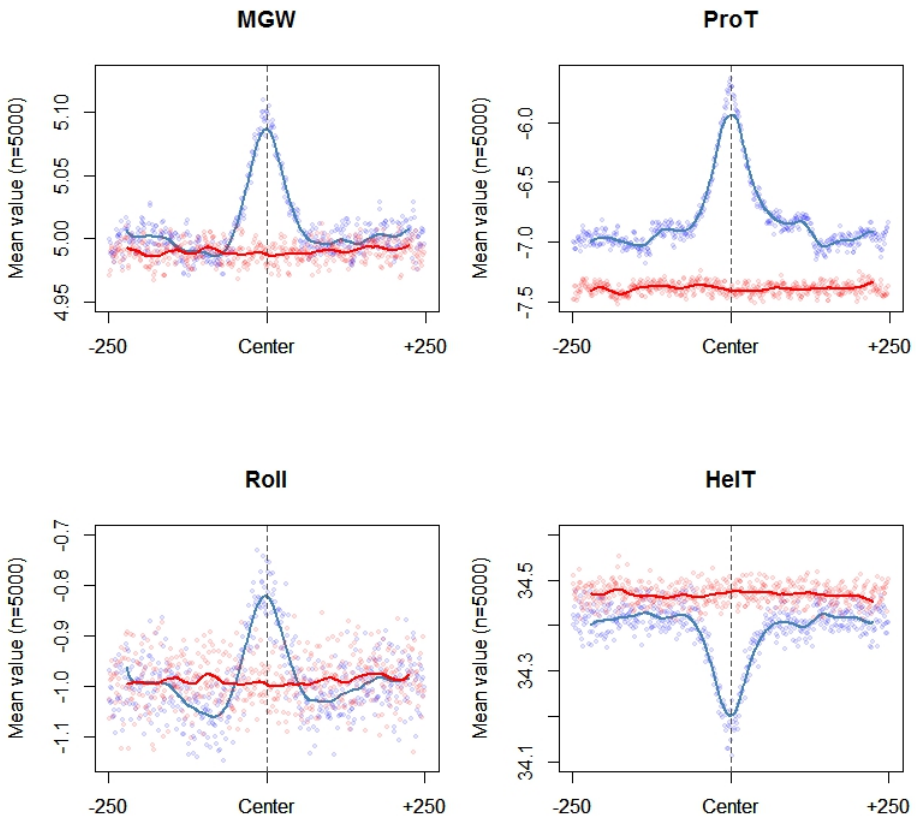

Figure 2: DNA shape parameters of minor groove width (MGW, Å), propeller twist(ProT,°), Roll,°, and helix twist(HelT,°) based on the sequence files of "bound 500.fa"(BLUE) and "unbound 500.fa"(RED).

Q8. Build prediction models for in vitro data

The shape-augmented model, "1-mer+shape", and the sequence-only model, "1-mer", perform equally well in distinguishing bound regions of DNA from unbound regions. The AUC scores for the ROC curves of the models, based on "bound 30.fa" and "unbound 30.fa" sequence files: $AUC_{sequence} \approx 0.84$, $AUC_{sequence+shape} \approx 0.842$, see Fig.3.

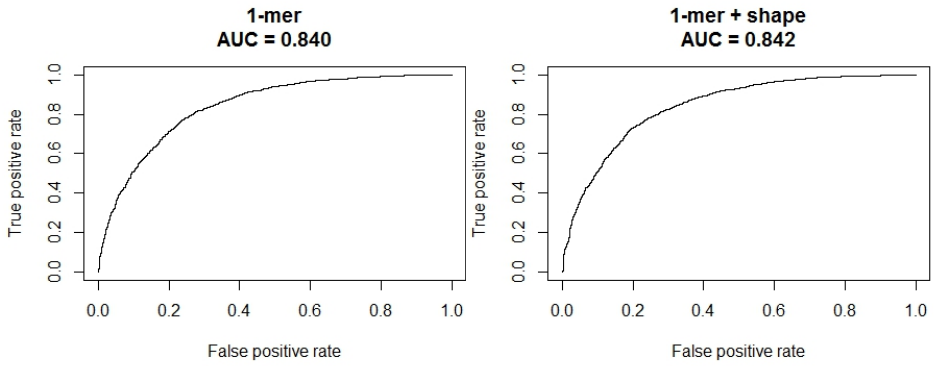

Figure 3: ROC for the classification models based on the "1 mer" sequence-only model and "1 mer + shape" shape-augmented model.

References

[1] Shamir R. Orenstein Y. *A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data.* Nucleic Acids Research, 2014.

[2] Liu P Abe N Slattery M1, Riley T. *Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins.* Cell, 2011.

[3] Wikipedia. *SELEX-seq.*