BISC-577 Module 3 Assignment 2

Guilherme de Sena Brandine

May 9, 2016

1 An open-source and distributed revision control project

The entire project was done on GitHub and can be found at:

https://github.com/guilhermesena/BISC577-HW2

The source code for each specific section of this report are on the root folder and called **question_x.r** where $x \in \{4, 5, 6, 7, 8\}$ and this report can be found in the **report** directory. All input files (fasta files or the ones generated by DNAshapeR) are located in the **in** folder except for the AH28451 CTCF binding site regions on question 6 due to its large file size.

2 High throughput binding assays

SELEX-Seq (Systematic Evolution of Ligands by Exponential Enrichment, aka in vitro selection): It is a method for characterizing the complete repertoire of binding site preference for transcription factor complexes. It consists of constructing a DNA library on an array and introducing the desired TF to the array, which allows us to quantify which k-mers the protein prefers to bind to, the major drawback being that we are limited to small k-mers (10 to 14 max) as the number of possible k-mers grows exponentially with the k-mer size.

PBM (Protein Binding Microarray): In terms of experimental design, PBM is the reverse of SELEX and measures the DNA affinity to transcription factors. This time proteins are the ones bound to an array and DNA probes are allowed to flow freely in the solution. If they bind to the array proteins, they can be later amplified by PCR and sequenced to quantify the k-mers that have binding specificity to the proteins in the array. The disadvantage of this method is that we can't do genome-wide studies of TF binding through this method as we need to have previous knowledge of what proteins to insert into the array.

ChIp-Seq: Unlike the former experiments, ChIp-Seq (Chromatin Immunoprecipitation followed by Sequenicng) is done in vivo. The idea is to use formaldehyde to cross-link proteins to the DNA and subsequently use antibodies specific to the desired protein (ie, TFs, histones, enhancers, etc) to immunoprecipitate the DNA-protein complex. We can then sequence the DNA regions that are IP-ed with the protein to find enriched regions in the genome.

3 Preparation of high throughput in vitro data analysis

I used the r portal to download R-3.3.0 for unix and the bioconductor source within R to download DNAshapeR through the biocLite() module. I also use the built-in install.packages() module to download caret from the R could mirror.

4 Build prediction models for in vitro data

The source code for this problem can be found at file **question_4.r**. I used caret to train a model through L2-regularization to predict the binding affinity of DNA sequences. I compared two cases: using only the 1-mer feature or the 1-mer+shape by using shape data from the DNA, and the makeSummary() function in the source code returns the maximum average R-squared for the simulation. These are the values I obtained for the 3 given input files:

	Mad	Max	Myc
1-mer	0.775	0.785	0.778
1-shape	0.863	0.864	0.855

Of course, the algorithm for model generation is randomized so the values we obtain for R^2 are slightly different each time we run it, but they always revolve around those values. We thus notice qualitatively that adding the shape feature increases the model performance.

5 High throughput in vitro data analysis

The plot for the above table is given in figure 1, and the code to built the plot and calculate the p-value is given in **question_5.r**.
Since our dataset consists of two paired samples which do not necessarily follow a normal distribution, the Wilcoxon matched pair test was used to calculate the p-value using the *coin* library's built-in wilcox.test() function. A p-value of 0.1 was found, and since it is higher than 0.05, we cannot discard the null hypothesis that the R^2 values for 1-mer and 1-mer+shape are uncorrelated.

6 Preparation of high throughput in vivo data analysis

The source **question_6.r** has the code used to download the AH28451 dataset from AnnotationHub into a file called **mmus.fa** that was saved inside the in directory (not shown in GitHub). We used width=400 to obtain 42,000 sequences of length 400.

7 High throughput in vivo data analysis

The plots given by DNAshapeR are given on figure 2. From the dataset we can infer that the center region of the CTCF binding site have larger minor groove width, rool and propeller twist and smaller helix twist. We can hypothesize that the minor groove opening and the smaller twist stress in the DNA double helix are important factors for the CTCF specificity. We also know that CTCF binds to the CCCTC motif, so it is possible that the motif itself may cause such behaviour in the DNA shape.

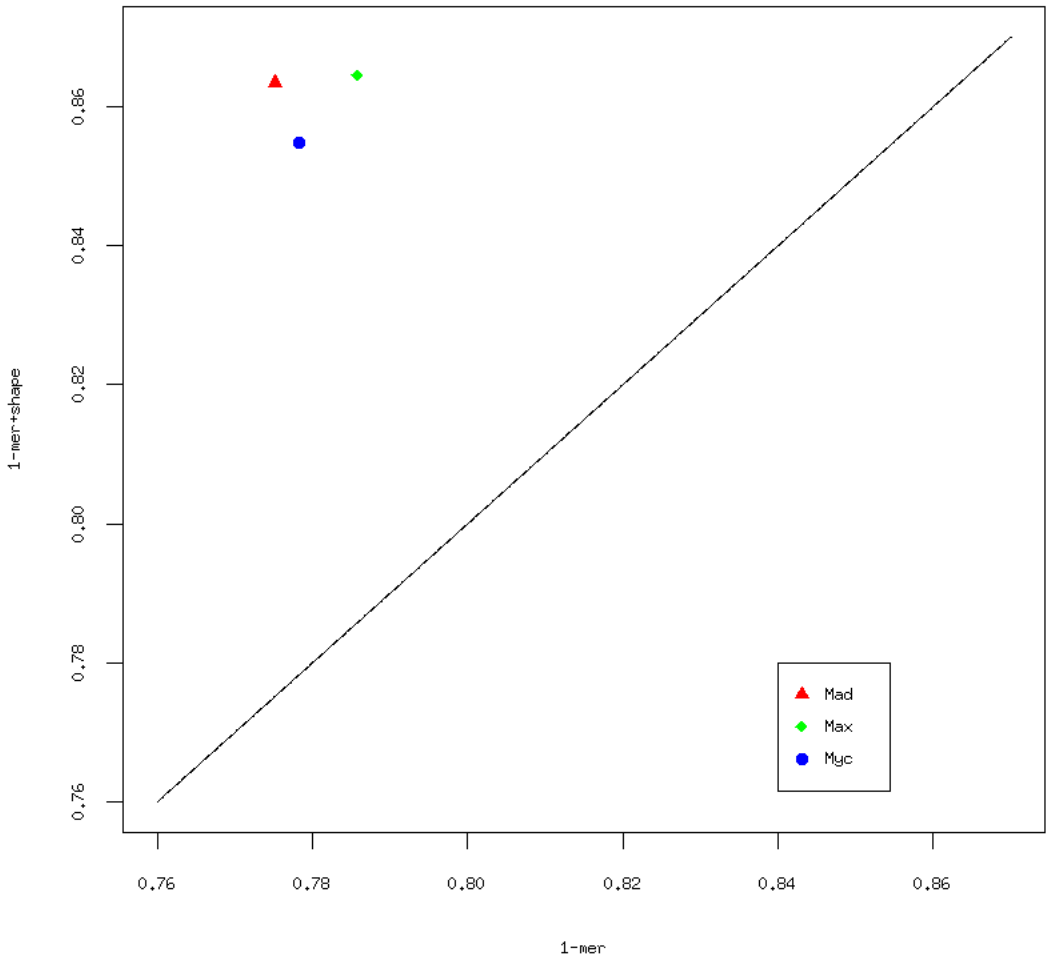

Figure 1: comparison of R^2 averages for each dataset.

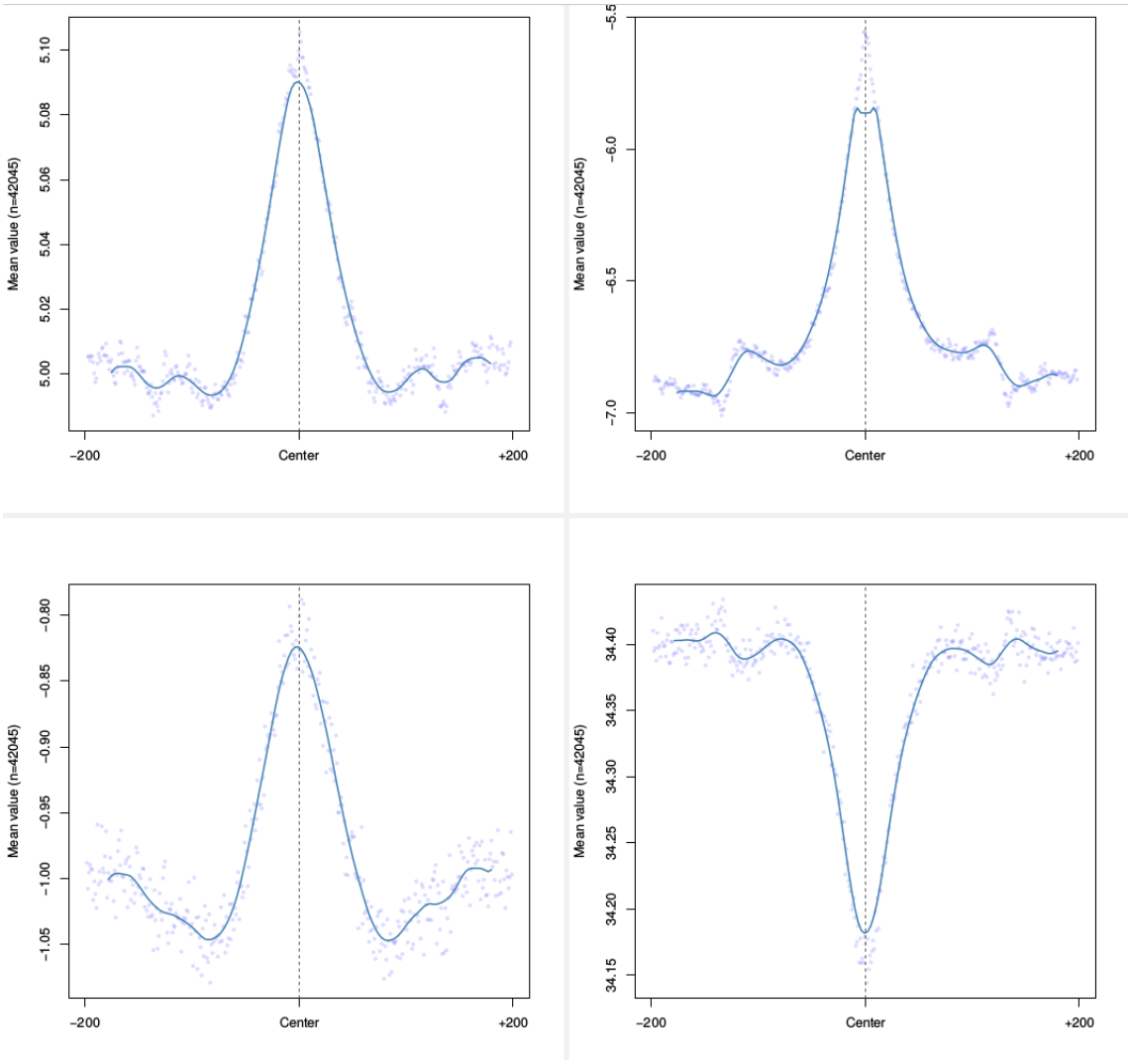

Figure 2: Minor Groove Width (top-left), propeller-twist(top-right), roll(bottom-left) and helix twist (bottom-right) for the CTCF transcription factor binding sites

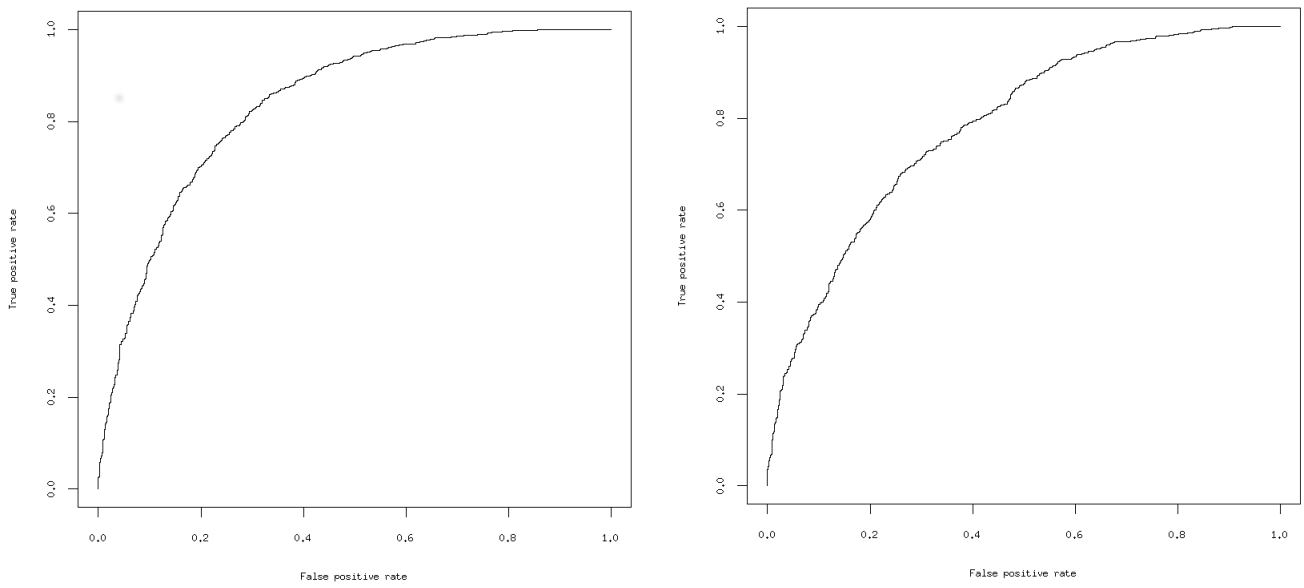

Figure 3: ROC curves for prediction model with 1-mer (left) and 1-mer+shape (right) features.

8 Build prediction models for in vivo data analysis

The code to generate non-overlapping regions and plot the ROC curve for the logistic regression model is in **question_8.r**. The idea is to create a new feature *isBound* that will be 1 for the dataset downloaded by AnnotationHub and 0 for the random manually generated sequences. Afterwards we train the model by adjusting a logistic function and a threshold and analyse the area under curve from the precision/recall plot to measure the prediction accuracy (AUC = 1 means a perfect model and AUC = 0.5 means a totally random guess). The two AUC plots are shown on Figure 3. For 1-mer we have an AUC of 0.835 and for 1-mer+shape we have an AUC of 0.782, from which we can infer that the sequence itself is sufficient to predict binding specificity and adding the shape feature decreases the prediction performance.