BISC 481, Fall 2016

Third Assignment for 1st section of the course (Rohs)

**Modeling of protein-DNA binding specificity/Statistical machine learning**

1. (a) An account was created on Github using my email address.
2. The username is Mkarim1 and the file name is: BISC-481-homework-3-Karim.
3. Two collaborators were added: TsuPeiChiu and xinbeibei
4. High-throughput binding assays:
5. In vitro experiments SELEX-seq and PBM: These experiments are performed in lab setting outside the body.

SELEX-seq stands for: Systematic Evolution of Ligands by Exponential enrichment). It is a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Importantly, it is flexible and can determine the relative affinities to any DNA sequence for any transcription factor or multiprotein complex. It explores the consensus sequence for nucleic acid binding protein through repeated cycles of binding selection and PCR amplification. For the procedure, the sequence library is derived from the chromosomal DNA of the target organism. PBM stands for Protein Binding Microarray. It involves using double stranded DNA microarray. An epitope –tagged TF is bound to dsDNA microarray and a tagged antibody to epitope is labeled with fluorophore. Then the microarray is scanned and the binding signal is recorded.

1. In vivo CHIP-seq: This is performed inside the body. It is a method used to analyze protein interactions with DNA. CHIP-seq combines chromatin immunoprecipitation with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.
2. The advantages of SELEX seq is that it is more flexible and there is no limit to the size of binding site. It can capture binding measurements of large protein complexes and quantity of binding is only limited by the depth of sequencing. While PBM can track large number of proteins in a parallel way. It is fast and relatively inexpensive. However, CHIP-seq advantage is that it increases resolution and reduces noise.

The disadvantages of SELEX seq is that it is relatively expensive and the initial pool contains strong sequence biases and additional biases are introduced during the many rounds of PCR amplification. Also, it provides integer-valued, Poisson-distributed sequence read counts. In PBM, the proteins are difficult to handle and arrays only cover 10 mer, making it difficult to model binding sites larger than 10 base pairs. On the other hand, CHIP-seq is relatively expensive and requires a lot of tissue.

1. Preparation of high-throughput *in vitro* data analysis:
2. R studio was downloaded for my windows laptop.
3. *Bioconductor* was installed from the file under name: BISC 481-master
4. *DNAshaper* package was installed using the provided code: biocLite("DNAshapeR")
5. Caret package was installed using the provided code: install.packages("caret")
6. The data information for *Mad*, *Max* and *Myc* were downloaded.
7. Build prediction models for in vitro data:
8. A feature vector was generated using *DNAshapeR* package using the codes provided in MLR\_example.R. To obtain data for Mad, we use the codes provided (some of which are):

fn\_fasta <- paste0(workingPath, "Mad.txt.fa")

pred <- getShape(fn\_fasta)

featureType <- c("1-mer", "1-shape")

-To obtain the data for Myc and Max, we replace Mad by Myc or Max. We use the same code to obtain the results for 1-mer, but we delete “1-shape” from the feature type.

1. L2-regularized MLR models were built using caret:

The code is obtained from the same file as DNAshapeR which is:

model2 <- train(affinity~., data = df, trControl=trainControl,

method = "glmnet", tuneGrid = data.frame(alpha = 0, lambda = c(2^c(-15:15))))

model2

result <- model2$results$Rsquared[1]

result

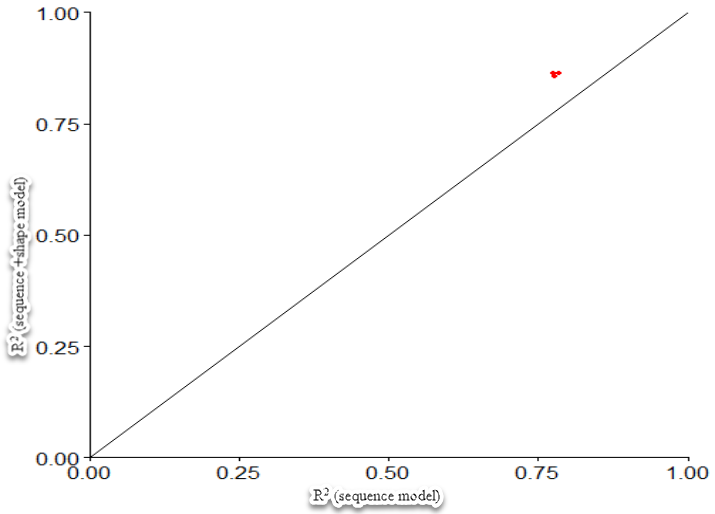
Then to obtain the data for Myc and Max, wherever there is Mad, it is replaced by Myc or Max.

R2 values are obtained using the codes provided. The results are presented in Table 1.

Table 1: R2 values for Mad, Max and Myc

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

1. High-throughput in vitro data analysis:
2. The codes from plotting\_example.R file were used and the only modification was that instead of two data points there were three. So the code (lines 28 and 29) was modified and the results obtained in Table 1 are input in these lines. The values of R2 were encoded where 1-mer was represented data 1 (x-axis) and 1-mer + 1-shape was represented by data 2 (y-axis). The codes are: data1 <- c(0.775, 0.778, 0.785), data2 <- c(0.863, 0.855, 0.864). The results are shown in Figure 1.



*Figure 1: R2 values*

1. From the results, we can see that the three models yielded a good R2 value since it fits the linear regression model. When we had sequence model alone, the values were close to 0.78, but when shape was included the value went up to 0.8 which is closer to 1 and thus more accurate. Moreover, if the point was on the diagonal, we wouldn’t had any gain, but in this case, the points are not on the diagonal and thus adding shape helped improve the values.
2. (a) CHIP-seq data was downloaded
3. R packages are installed
4. (a) The codes were used to plot the three parameters. The code was changed in line 17 (plotShape(pred$MGM)) where instead of MGW, we put ProT to get propeller twist and then HelT to get helix twist. The plots for bound are shown in Figures 2a, b, and c.

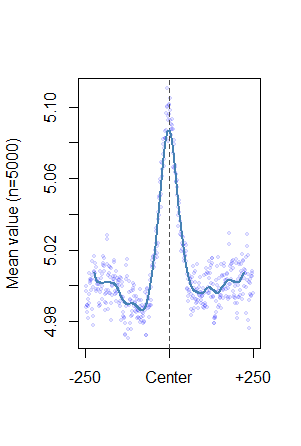
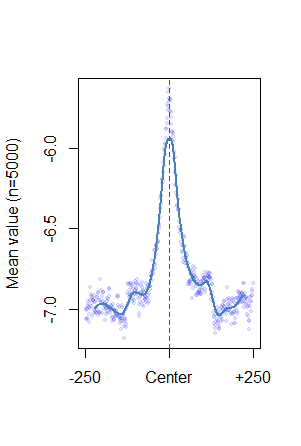


Figure 2b: ProT

Figure 2a: MGW

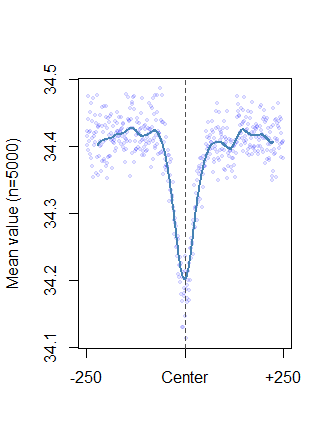


Figure 2c: HelT

1. These results can help predict the structure of DNA. For example, the Minor groove width in bound data is wider than unbound, and this is the shows how the transcription factor might recognize it. HelT and ProT plots can be used and compared as well to learn how they act for the active site.
2. Build prediction models for in vitro data:
3. Auc value for “1-mer, 1-shape” = 0.836 and the graph shape is shown in figure 3a below.

Auc value for “1-mer”= 0.839, and the graph is shown in Figure 3b. These values were obtained using the codes:

## Plot AUROC

prediction <- prediction( model$pred$Y, model$pred$obs )

performance <- performance( prediction, "tpr", "fpr" )

plot(performance)

## Caluculate AUROC

auc <- performance(prediction, "auc")

auc <- unlist(slot(auc, "y.values"))

auc

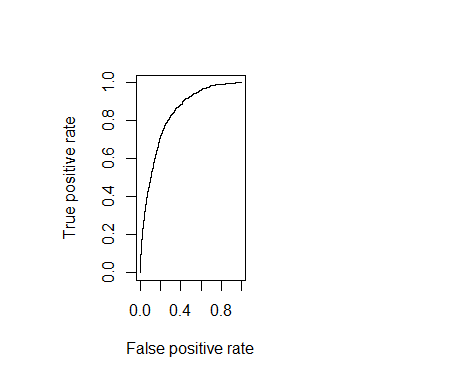
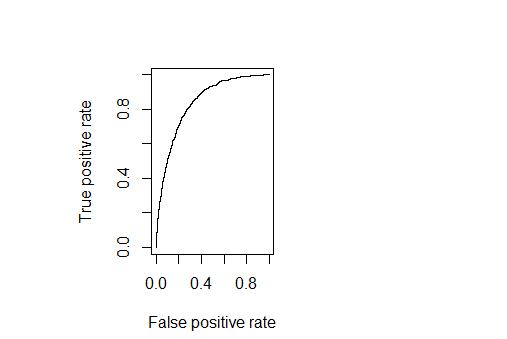


Figure 3a: AUC plot for 1-mer

Figure b: AUC plot for 1-mer+1-shape

1. The values of Auc obtained are similar. The values are close to 1, which means that the results are good but not perfect. Since the values are similar, the shape of the DNA may not be the way to study this DNA. Question 4 gave us different values and improved the values when we introduced shape.