Assignment 8

Due Wednesday, October 28 before midnight (California time)

We are going to build models to predict protein–DNA binding based on the data obtained from high-throughput *in vitro* protein–DNA binding assays.

Download and install the following packages from Jupyter Notebook or other platforms:

a. Bioconductor (refer the instructions on https://www.bioconductor.org/install/)
b. The DNA shape prediction tool: DNAshapeR (refer the instructions on https://www.bioconductor.org/packages/release/bioc/html/DNAshapeR.html)
c. The machine learning tool: caret (refer the instructions on https://github.com/topepo/caret)

Complete the following tasks:

1. Use the DNAshapeR package to predict DNA shape for each sequence bound by transcription factor *Mad, Max and Myc*. The datasets were obtained from the *in vitro* gcPBM (genomic context protein binding microarray) assay. The sequence data in FASTA format (Mad.fa, Max.fa and Myc.fa) can be found on Blackboard. Use plotShape() or heatShape() functions of DNAshapeR to generate ensemble plots for the DNA shape parameters of minor groove width (MGW), propeller twist (ProT), Roll, and helix twist (HelT). Discuss what you find from the results [2pt].

2. Write an R function that takes as input a FASTA file. The function returns a feature matrix for "1-mer" sequence model using one-hot encoding which is a way to represent categorical data as binary vector [3pt].
   For DNA, we have four categories A, T, G, and C. Thus a one hot code for DNA is:
   A: [0,0,0,1]
   C: [0,0,1,0]
   G: [0,1,0,0]
   T: [1,0,0,0]

   For example, for the sequence AATTC, the 1-mer one-hot encoding is:
   [0,0,0,1,0,0,0,1,1,0,0,0,1,0,0,0,1,0,0,0]

3. Use the function from Q#2 to generate a "1-mer" feature vector and combine the shape features generated from Q#1 for "1-mer+shape" model with respect to the datasets of *Mad*, *Max* and *Myc*. Use the caret package to build L2-regularized MLR models for "1-mer" and "1-mer+shape" features with 10-fold cross validation (set the L2 *lambda* value between 2^{-15} to 2^{15}), and print out the average R2 (coefficient of determination) for these two models with respect to

the datasets of *Mad*, *Max* and *Myc*. The corresponding protein–DNA binding affinities can be found in corresponding *.s files. Compare the performance between “1-mer” sequence model and “1-mer+shape” model and explain what you find [3pt].

Turn in the code for the aforementioned R functions and the answers into one file in Jupyter Notebook format (.ipynb). Use the “Turnitin” link on Blackboard/Assignments/Assignment 8 to submit this file.