**Task: to predict potential RBP binding sites in SARS-CoV-2.**

RNA binding proteins (RBPs) are proteins with various functions which can binding to RNAs. They usually bind to specific regions of RNAs by recognizing local sequence features. The goal of this homework is to find whether and where they can bind to the SARS-CoV-2 (2019-nCoV) RNA sequences.

1. Please build your model for predicting RBP binding sites. We provide positive and negative samples (i.e., binding and unbinding sequences) of four RBPs, in files with suffix “train.positives.fa” and “train.negatives.fa”. Please choose 4 to 5 RBPs, and train a model for each of them. Building a prediction model for MOV10 with basic k-mer features and logistic regression as an example can be found in our sample code. You are encouraged to use alternative ways to extract the sequence features and build the prediction models, e.g., one-hot encoding features and CNN/RNN models. Please describe your method and report the AUC scores of the validation process (calculated in the “training” function).
2. Please use your trained model to predict potential RBP binding sites on SARS-CoV-2 sequences. You can provide the genomic coordinates and sequence contents of the predicted binding sites. Alternatively, sequence motifs can be used to illustrate the results. Motifs can be generated by online tools (e.g., <http://weblogo.berkeley.edu/logo.cgi>) or by your algorithms.

**References:**

Data source:

Maticzka,D., Lange,S.J., Costa,F. and Backofen,R. (2014)

GraphProt: modeling binding preferences of RNA-binding proteins.

Genome Biol., 15, R17.

Feature extraction and a machine learning model:

Li, S., Dong, F., Wu, Y., Zhang, S., Zhang, C., Liu, X., ... & Zeng, J. (2017).

A deep boosting based approach for capturing the sequence binding preferences of RNA-binding proteins from high-throughput CLIP-seq data.

Nucleic acids research, 45(14), e129-e129.

A neural network model:

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), 831-838.