QBIO 481, Fall 2023

Optional Assignment 100 pts.

1. Application of an open-source and distributed version control project: (a) Create a public repository with a README at GitHub <https://github.com>. (b) Write your name in the file of README.md. (c) **You are required to push your report and R scripts to the repository.** The example and file templates are shown in <https://github.com/TsuPeiChiu/QBIO481>. 5 pts.
2. High-throughput binding assays: Briefly describe (a) the *in vitro* experiments SELEX-seq and PBM, and (b) the *in vivo* experiment ChIP-seq. (c) Compare and discuss the advantage and disadvantage of these methods. 10 pts.
3. Preparation of high-throughput *in vitro* data analysis: (a) Download and install R (version >= 3.3.0) from <https://www.r-project.org>. (b) Install *Bioconductor* on your R platform. The installing instruction can be found at <https://www.bioconductor.org/install/>. (c) Install package *DNAshapeR* on your R platform. The installing instruction can be found at <https://www.bioconductor.org/packages/devel/bioc/html/DNAshapeR.html> (d) Install the machine learning package *caret* on your R platform. The installing instruction can be found at <https://github.com/topepo/caret> (e) Download the gcPBM *in vitro* experimental data of *Mad, Max and Myc* from <https://github.com/TsuPeiChiu/QBIO481/tree/master/gcPBM>. 10 pts.
4. Build prediction models for *in vitro* data: (a) Use the *DNAshapeR* package to generate a feature vector for “1-mer” sequence model and a feature vector for “1-mer+shape” model with respect to the datasets of *Mad, Max and Myc*. (b) Use the *caret* package to build L2-regularized MLR models for “1-mer” and “1-mer+shape” features with 10-fold cross validation, and print out the average R2 (coefficient of determination) for these two models with respect to the datasets of *Mad, Max* and *Myc*. 20 pts.
5. High-throughput *in vitro* data analysis: (a) Draw a plot for a comparison of two different models (1mer vs. 1mer+shape) as shown in Figure 1(B) of Zhou et al. PNAS 2015 (b) Briefly discuss what you have learned from the results. 15 pts.
6. Preparation of high-throughput *in vivo* data analysis: (a) Download the ChIP-seq data (including “bound” and “non-bound” data) of CTCF transcription factor of *Mus musculus* from <https://github.com/TsuPeiChiu/QBIO481/tree/master/CTCF>. (b) Install the R packages mentioned in question (3). 5 pts.
7. High-throughput *in vivo* data analysis: (a) 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) based on the sequences downloaded for question (6). (b) Briefly discuss what you have learned from the results. 15 pts.
8. Build prediction models for *in vitro* data: (a) Build logistic regression models for “1-mer” and “1-mer+shape” features, draw a plot of the ROC curves, and calculate the AUC score for each curve. (b) Briefly discuss what you have learned from the results. 20 pts.

\*\* Please clearly comment your code. Points would be deducted otherwise.

Contact Raktim Mitra ([raktimmi@usc.edu](mailto:raktimmi@usc.edu); RRI 413M) or Professor Rohs ([rohs@usc.edu](mailto:rohs@usc.edu); RRI 413H) for help if needed.

Submission Instructions: Submit a link to your GitHub repository through blackboard.

Due: Tuesday, October 24, 2023