Justin Lubin, Colin Skinner

April 10, 2023

CS 289A Project Proposal

Building a model for inference of CRISPRoff efficacy based on gene characteristics

1. Background information

The introduction of CRISPR revolutionized the world of biology by giving us the capability to quickly and easily modify the genome of an organism, with applications from agriculture to therapeutics. However, CRISPR is difficult to reverse and toxic to cells, making it difficult to administer for human therapeutic purposes.

CRISPRoff is a molecule developed by Nuñez et al. in 2021 [1] that, like CRISPR, can suppress gene expression—however, it does so without modifying the genome, resulting in easy reversibility and less toxicity to cells, making it a prime candidate for therapeutic development. It works by modifying the molecules that attach to the genome—the so-called “epigenome”—but the exact mechanism by which CRISPRoff works is still unknown.

In particular, CRISPRoff works really well when targeting some genes, but not others—and biologists do not fully understand why. In this project, we aim to make progress on understanding what features of genes make them good targets for CRISPRoff. Specifically, we will build a model for inference of CRISPRoff efficacy based on gene characteristics.

2. Data sources

We have aggregated data from ten different sources to form a design matrix with 17,415 samples and 11 features. Each sample point is a gene, and each of the 11 features correspond to an attribute of that gene determined experimentally, such as the length of the gene, how highly-expressed the gene is, the percentage of certain chemical modifications found on the genome at that particular gene, and so on. For each of the 17,415 genes, we also have a CRISPRoff “score” that indicates how well CRISPRoff works on that particular gene; our goal will thus be to predict this score from our design matrix in a manner that supports scientific inference.

3. Methods

We will train and optimize a LASSO regression model to predict CRISPRoff efficacy from genetic features. We have efficacy data in the form of experimentally-derived, continuous-valued scores so our task is therefore a regression problem. We have $n \gg d$, so a simple least-squares regression model *could* be made without the need for regularization. However, a secondary goal is to infer whether certain genetic features are more critical in predicting CRISPRoff efficacy so that researchers could direct their focus on those features when trying to determine a molecular mechanism for CRISPRoff. Therefore, the feature selection afforded by L1 penalization is why we are choosing to build a Lasso regression model.

We have both numerical and categorical features, so we will use one-hot encoding for the categorical features. We will perform a random split of our data into training and test sets. Our numerical features have different units and ranges so we will scale the data, testing both standardization and normalization to see which yields better model performance. To prevent data leakage, we will fit the scaling parameters to the training data before transforming it with those parameters, and we will transform the test data with the training set scaling parameters. During training, cross-validation will be used to tune the L1 penalty hyperparameter. We will use several measures of model performance: the mean squared error, ROC-AUC and median absolute error. Cross-validation will be used to find the optimum performance on the validation set (a subset of the training data) and the final model will be evaluated based on performance on the test data.

4. Preliminary work

The bulk of the work that we have done thus far is combining and cleaning the data from many different sources. (This has been challenging, but the design matrix is completely pre-processed and ready-to-analyze now.) We have not yet tried any machine learning models on the data.

5. Project core

The core work of our project will be the optimization of the model, which will primarily involve an exploration of methods and the data.

References

[1] Nuñez, J.K., Chen, J., Pommier, G.C., Cogan, J.Z., Replogle, J.M., Adriaens, C., Ramadoss, G.N., Shi, Q., Hung, K.L., Samelson, A.J., Pogson, A.N., Kim, J.Y.S., Chung, A., Leonetti, M.D., Chang, H.Y., Kampmann, M., Bernstein, B.E., Hovestadt, V., Gilbert, L.A., Weissman, J.S. “Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing.” *Cell* 184 (2021).