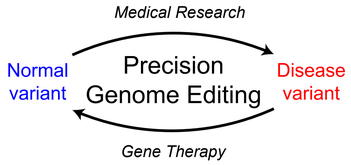
Optimizing Precision Genome Editing through Machine Learning

Shi-An A. Chen[1] and Elizabeth Tran[2, 3]

[1]Department of Biology, [2]Department of Biomedical Informatics, [3] Department of Psychiatry and Behavioral Science

BACKGROUND



Precision genome editing allows us to install various sequence into the genome and assay their effects on phenotype. Once we gain knowledge regarding the genetic cause of the disease, it is possible to use precision genome editing again to correct the mutated DNA to that of a healthy person.

CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9 is a popular gene editing system that introduces precise cuts in the genome by programming molecules called guide RNAs. The genome can be precisely edited by supplying "donor DNA" after cleavage by guide RNA-programmed Cas9, with which the cell can use as template for filling in lost DNA and thus incorporating sequences provided by the donor DNA, a process know as homology-directed repair (HDR).

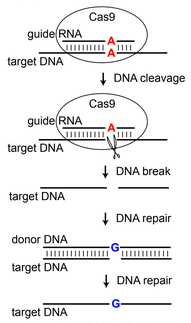


Figure 1. Example of CRISPR editing

Is there a better way?

Can we identify suboptimal guide RNA/donor DNA pairs beforehand and exclude them from experimental design?

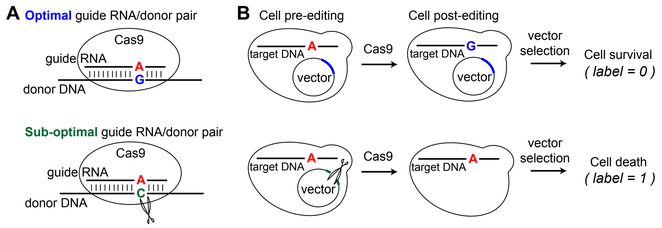
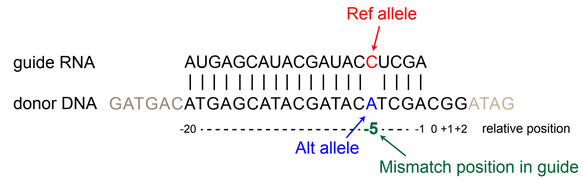


Figure 3. Optimal vs. Suboptimal guide RNA/donor DNA pairs during CRISPEY editing.

CRISPEY (Cas9-Retron precISe Parallel Editing via homologY), a novel genome-editing technique developed at Stanford in 2018, empirically measures in parallel the fitness effects of thousands of natural genetic variants in yeast at single-base resolution. However, co-introduction of donor DNA and guide RNA on the same vector meant that suboptimal guide RNA/donor DNA pairs will lead to cutting of the donor DNA by Cas9, which is observed in the editing phase of CRISPEY experiments (Fig. 3A). Such guide RNA/donor DNAs lead to self-destruction of editing vector and subsequently cell death, providing no information about the edit and waste of experimental throughput (Fig. 3B).

METHODS

We use the CRISPEY dataset from Sharon et al to extract guide RNA features as input and off-target effect as labels, which is a rich resource for modeling off-target effect with thousands of guide RNAs. The dataset consisted of 18,719 samples, 249 of those samples are labeled as 1 (cell death during editing phase) while 18,468 of those are labeled as 0 (survival during editing phase).



Since cutting of DNA requires perfect match between guide RNA and donor DNA, we pre-processed the raw guide RNA and donor DNA sequence into 3 features and use them as input: mismatched base in the guide RNA; mismatched base in the donor DNA and the position of mismatch between the guide RNA and donor DNA as features.

EXPERIMENTS

With these three features, we explored the performance of traditional machine-learning algorithms (i.e. Support Vector Machine, Logistic Regression, Random Forest) with various hyperparameters (different forms of kernels, Cs, and tree estimators) to a simple deep neural network (DNN).

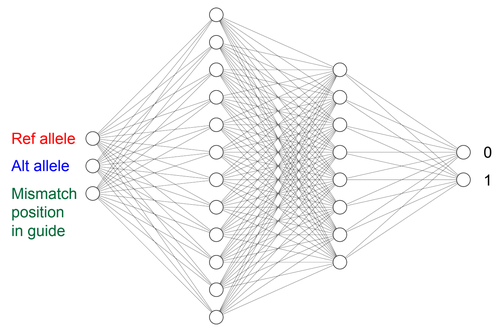


Figure 3. Simple Deep Neural Network (DNN) Architecture: The input layer is activated with ReLu function and followed by a dropout layer with a rate of 0.25. The first hidden layer is also activated by a ReLu function, but is followed by a batch normalization layer and a dropout layer of .10. The next hidden layer is activated by a softmax function with a cross-entropy loss.

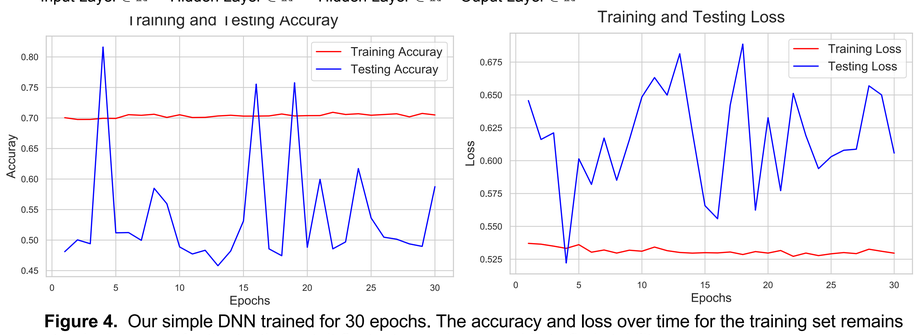


Figure 4. Our simple DNN trained for 30 epochs. The accuracy and loss over time for the training set remains steady; however, for the testing set, it fluctuates and was unable to find stability.

RESULTS

Model Type	Accuracy	Recall	Precision
Logistic Regression (C= 0.01)	**94 %**	8.77 %	**2.20**
Support Vector Machine (Kenral RBF, C = 1)	30.76 %	**64.15 %**	1.10 %
Random Decision Forest (C=0.01, l2)	**85.40 %**	15.78 %	1.39 %
Neural Network I (* check parameters on Table X)	**62 %**	37 %	1.03 %

Table 1. Comparison of our best performing model metrics.

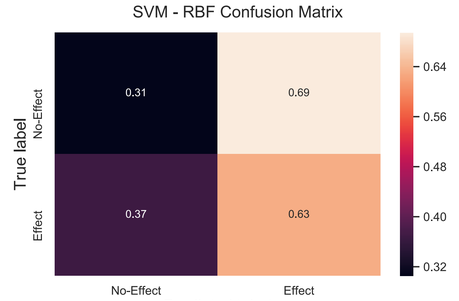


Figure 5. We can observe that there is a 69% false positives on the effect class resulting in a type I error. Though we have implemented SMOTE to address our imbalance classes, our new data points may be sampled from several our extreme values.

CONCLUSION

Different methods currently exist to detect CRISPR off-target mutations; however, they come with limitations and thus need to be identified experimentally. Having a reliable machine learning model to make prediction suboptimal guide RNA and associated donor DNA pairs can contribute to better identifying of off-target edits in precision editing studies. We compared the performance of three traditional machine-learning method algorithm with a DNN. Our SVM model is the highest recall performing model with a rate of 64%. Our logistic regression model is the highest accuracy performing model with a rate of 94%. The SVM model that we implemented is at a performance where we believe could be used for precision genome editing.

NEXT STEPS

- Add in additional features, such as guide RNA sequence, cutting efficiency, DNA shape, etc.
- Explore our DNN with different hyperparameters (e.g., regularization, dropout rate, learning rate)

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Sharon, E., Chen, S. A. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., & Fraser, H. B. (2018). Functional genetic variants revealed by massively parallel precise genome editing. Cell, 175(2), 544-557