Optimizing Precision Genome Editing through Machine Learning

Shi-An A. Chen¹ and Elizabeth Tran², ³

¹Department of Biology, ²Department of Biomedical Informatics, ³Department 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 programmable 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 known as homology-directed repair (HDR).

CRISPEY (Cas9-Retron preCrISe Parallel Editing via homologyY), 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 DNA 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).

RESULTS


<table>
<thead>
<tr>
<th>Model Type</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>94 %</td>
<td>8.77 %</td>
<td>2.20</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>30.76 %</td>
<td>64.15 %</td>
<td>1.10 %</td>
</tr>
<tr>
<td>Random Decision Forest</td>
<td>85.40 %</td>
<td>15.78 %</td>
<td>1.39 %</td>
</tr>
<tr>
<td>Neural Network I</td>
<td>62 %</td>
<td>37 %</td>
<td>1.03 %</td>
</tr>
</tbody>
</table>

Table 1. Comparison of our best performing model metrics.

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 84%. 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:
