SINGLE GUIDE RNA (sgRNA) EFFICIENCY PREDICTION

BELLO, KAZEEM ENITAN LinkedIn: www.linkedin.com/in/kazeembello

Introduction

The CRISPR/Cas9 system has revolutionized genetic research and biotechnology, offering precise and efficient tools for genome editing. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and Cas9 (CRISPR-associated protein 9) are part of a natural defence mechanism found in bacteria and archaea, which they use to fend off viral attacks. When a virus invades a host organism, a piece of the viral DNA is integrated into the CRISPR locus in the form of "spacers." This acts as a molecular memory of past infections. Subsequently, when the same virus attacks again, the CRISPR locus is transcribed into CRISPR RNA (crRNA) and combined with trans-activating crRNA (tracrRNA) [3]. Together, they form a single guide RNA (sgRNA) which guides the Cas9 protein to the precise target site in the invading viral DNA. Cas9 acts as molecular scissors, cleaving the DNA at the specified location as shown in fig. 1. However, not all sgRNAs exhibit the same efficiency in gene targeting, leading to the need for a reliable optimization method to identify the most efficient guides [2].

To effectively apply the CRISPR-Cas9 system for gene editing, researchers need to identify target sites that can be cleaved efficiently (on-target prediction) and for which the candidate gRNAs have little or no cleavage at other genomic locations (off-target prediction)[1]. Developing a machine learning model to predict the on-target and off-target cutting efficacies is a critical component of effective sgRNA design for CRISPR experiments.

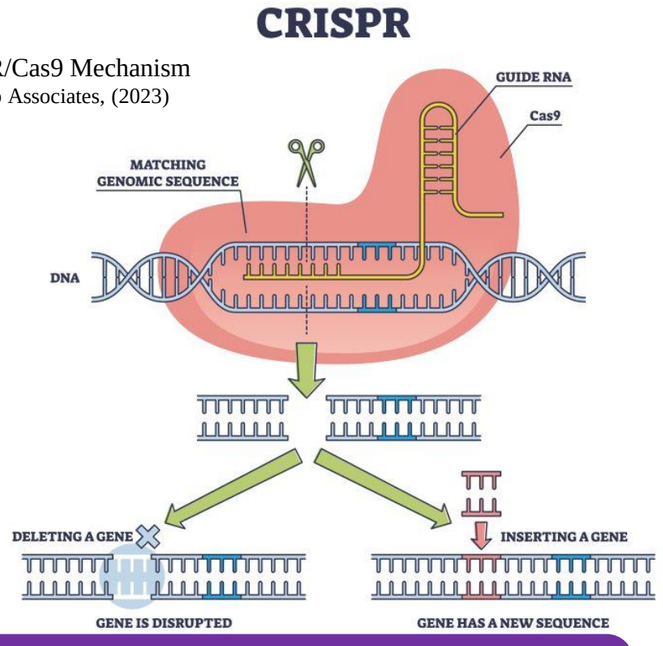

Fig. 1 CRISPR/Cas9 Mechanism
Source: Lab Associates, (2023)

Why is this important to CRUK?

- If we are more confident in the quality of a gRNA sequence, we can design more efficient libraries
- The sooner we can create a sgRNA library, the more efficiently we can process screens
- The better the efficiency of the sgRNA library we use, we can trust the results and better understand the effects of therapies on targets.
- The need to develop a model trained on CRUK in-house generated data.

Methodology

The choice of the feature used for training was based on literatures in this domain that have used similar feature and have been found to be effective in guide design. A 23base pair (bp) sequence composed of 20bp followed by 3bp Protospacer Adjacent Motif (PAM) in the form of $N_{20}NGG$ was used. N represent any nucleotides (Adenine (A), Cytosine (C), Guanine (G) and Thymine (T))

The target variable (efficiency) is an experimentally determined log fold changes. These were calculated at different days and the average was determined. The average fold change has a linear relationship with the fold changes of individual days. Hence, it was used as the target variable.

For model development, the 23bp sequence was encoded into a one-hot matrix of 4 x 23. This can also be thought of as a 1 channel (black and white) image of 4pixel by 23pixel (fig. 2).

4 = number of rows (A, C, G & T) 23 = number of columns (23bp sequence)

This matrix was inputted into a convolution neural network (CNN) which was able to learn feature representation using the convolution and pooling layer with RELU activation. The result of this is passed into a fully connected layer which is a regression layer that outputs the predicted value of the guide efficiency (fig. 3). The model was trained to predict only the on-target efficiency.

	A	G	C	T	A	G	A	C	A	C	A	T	C	C	G	A	T	T	G	A	C	C	G	G
A	1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
C	0	0	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0
G	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
T	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0

Fig. 2. Sequence Encoding

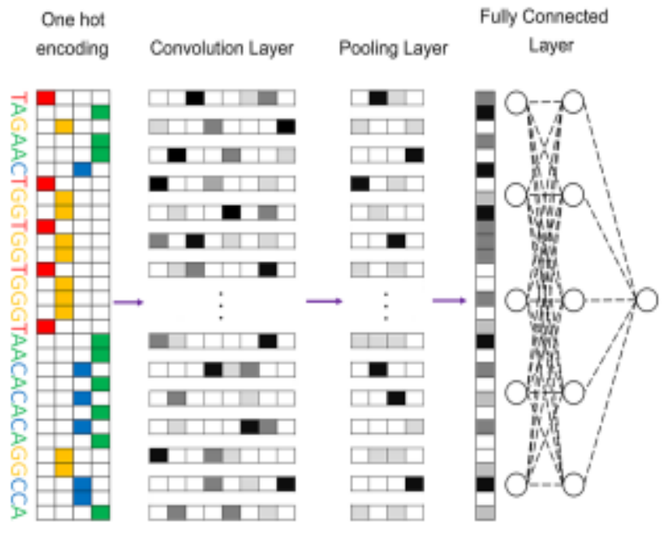

Fig. 3. Model Architecture (Xue et. al (2019))

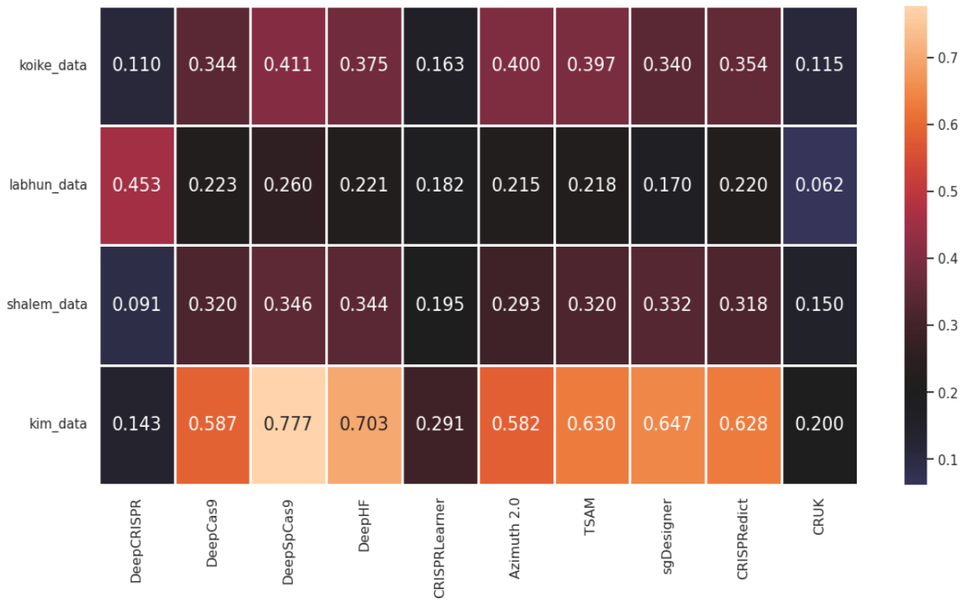

Fig. 4. Comparison of Models Using Spearman Correlation

Result

- We assessed the model's performance with R2-score and compared it to existing models using Spearman correlation.
- The initial model trained achieved a very low R^2-score of -0.2. After data augmentation, the model performance improved significantly when tested with CRUK sgRNA testing data with an R^2-score of 0.98. However, the model perform very poor when tested on four datasets from obtained previous studies achieving an R^2-score of -0.35, -0.9, -0.67 and -0.062 .
- Similarly, the spearman correlation of this model and nine other state of the art models on the four dataset was calculated and presented in the fig. 4. As evident the model, named CRUK is one of the less performing models with an average spearman correlation of 0.13 across the four datasets.

Discussion and Further Work

As seen from the result, the model did not generalize well and it overfits leading to a poor performance on dataset that were generated on different conditions of its training data (CRUK data). This is not the conclusion of the study. As this is a base model, work is going on to improve the performance of the model by:

- Optimization of the model via hyperparameter tuning to avoid overfitting and improve generalization.
- Using a 30bp sequence as a feature variable and also combining the sequence with other guide features such as thermodynamic features. The specific features that determine on-target activity remain largely unexplored [1]. Sometimes the poor performance of a model resides in the training data.
- Employing transfer learning by training on AWS custom models. These models have been trained on a large amount of dataset. A more generalized model is expected when our data is trained on these models.

Acknowledgement

- This machine learning project was conducted in Cancer Research UK, a registered charity in England and Wales. We want to bring about a world where everybody can lead longer, better lives, free from the fear of cancer.
- Special Thanks to:
  - Cancer Research UK: Technology, FGC Infrastructure, and Senior Leadership Teams
  - (CRUK) Cancer Research Horizons: Bioinformatics and Senior Leadership Teams
  - Cancer Research Horizons-AstraZeneca Functional Genomics Centre

References

1. Dimauro, G., Colagrande, P., Carlucci, R., Ventura, M., Bevilacqua, V. and Caivano, D., (2019). CRISPRLearner: A Deep Learning-Based System to Predict CRISPR/Cas9 sgRNA On-Target Cleavage Efficiency. Electronics, 8(12), p.1478.
2. Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J. and Root, D.E., (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature biotechnology, 32(12), pp.1262-1267.
3. Lab Associates, (2023). CRISPR: A Gene Editing Tool [online image] https://labassociates.com/crispr-a-gene-editing-too
4. Xue, L., Tang, B., Chen, W. and Luo, J., (2019) . Prediction of CRISPR sgRNA activity using a deep convolutional neural network. Journal of Chemical Information and Modeling, 59(1), pp.615-624