AI-Powered Prediction of Off-Target Effects in Genetic Engineering: A Bioinformatics and Machine Learning Approach

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Natalie Nieves

Spring, 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Briana Morrison, Department of Computer Science

Rosanna Vrugtman, Department of Engineering and Society

AI-Powered Prediction of Off-Target Effects in Genetic Engineering: A Bioinformatics and Machine Learning Approach

CS4991 Capstone Report, 2024

Natalie Nieves
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
Ncn7tvy@virginia.edu

ABSTRACT

The increasing concern over unintentional consequences in genetic engineering prompts the need for computational tools to predict and mitigate off-target effects. To lessen the possibility of unforeseen outcomes, I propose an Artificial Intelligence (AI) model designed to predict potential off-target effects of genetic modifications. The model would utilize advanced machine learning algorithms and use large and diverse genomic datasets to enhance accuracy. The design and implementation would involve the integration of deep learning algorithms and bioinformatics tools, requiring knowledge and skill in both computer science and genetics. Initial results indicate promising capabilities in identifying and understanding unintended consequences. However, further testing and evaluation are necessary to validate the model's accuracy across diverse genetic contexts.

1. INTRODUCTION

For decades, genetic engineering has impacted the field of biotechnology, allowing the modification of organisms in a range of applications, such as medicine, agriculture, and human biology While the advancements in genetic engineering have been promising, they have also brought many challenges. One major concern is the increasing likelihood of unintentional genetic changes, known as off-target effects. This occurs when gene modification tools unintentionally alter genetic sequences other than the target sequences. These unintentional modifications can lead to disruption in gene sequences, causing unpredictable outcomes.

Considering the growing concern about unintentional consequences, there is a need for advanced computational tools to predict off-target effects. To address this concern, this project introduces an AI model designed to predict these potential off-target effects in genetic modifications. By utilizing machine learning algorithms and diverse genomic datasets, the model aims to improve the accuracy of identifying off-target sites.

2. RELATED WORKS

The advancement of genetic engineering techniques has led to the development of various off-target prediction algorithms. According to Lin and Wong (2018), such algorithms include the CFD score, MIT score, and CROP-IT score; and these algorithms calculate scores based on mismatches to the guide RNA sequence, with higher scores indicating a higher likelihood of off-target activity.

According to the Genetic Perturbation Platform (2016): “The Cutting Frequency Determination (CFD) score is calculated by using the percent activity values provided in a matrix of penalties based on mismatches of each possible type at each position within the guide RNA sequence.” The MIT score aims to predict the off-target cutting rate by utilizing a mismatch weight matrix based on comprehensive research of gRNA variations and adjusts the final score based on the closest distance between the mismatches. (Bao, et al., 2020). Last, according to Singh, et. al. (2021): “The CROP-IT algorithm is based on a computational model where each position of the guiding RNA sequence is differentially weighted based on experimental Cas9 binding and cleavage site information from multiple independent sources.”

These algorithms offer significant advancements in predicting off-target effects. Although they have provided valuable insights and predictions, there remains a need for improved accuracy and reliability across various genetic contexts. Addressing this gap, the proposed project aims to develop an AI-powered algorithm designed to enhance these qualities.

3. PROJECT DESIGN

The main goal of this project design is to develop an Artificial Intelligence (AI)-powered algorithm specifically designed to have the ability to predict potential off-target effects in genetic modifications. This algorithm aims to lessen the possibility of non-deterministic outcomes in gene therapy by accurately identifying potential off-target sites to assist genetic engineers in making informed decisions during the implementation of genetic modifications**.**

3.1 Overview of AI Model

This AI algorithm will be implemented by using machine learning algorithms with bioinformatic tools. The machine learning portion of the algorithm would be a key tool in processing and analyzing given genomic datasets, while the bioinformatic tools will provide necessary genetic context to overall increase the accuracy of the predictions. It will use deep learning algorithms to analyze genomic sequences and predict off-target effects. This would involve the preprocessing of datasets and extraction of information from these datasets to use to make predictions based on patterns learned when training the AI model.

The model will be implemented for use across various genetic engineering techniques or contexts with the purpose of serving as a preliminary screening tool to help engineers determine potential off-target sites for future experiments. It is not intended to replace experimental validation methods but to provide feedback of potential off-target sites to be examined.

3.2 Data Collection and Preprocessing

For this project, I will use two methods of machine learning; supervised and unsupervised learning. For unsupervised learning, I will utilize the “CRISPRseek” dataset. This dataset consists of columns including "name," "gRNAPlusPAM," "targetInSeq1," "targetInSeq2," "scoreForSeq1," "scoreForSeq2," "mismatch.distance2PAM," "n.mismatch," "scoreDiff," and "targetSeqName.” To preprocess and prepare this data for the machine learning algorithms, I will use one-hot encoding for the column "targetSeqName" to change it into a format suitable for machine learning algorithms, clean the data in cases of null values, and extract essential data needed to train the AI model effectively. This will ensure that the AI model can effectively learn and train the data, which is crucial for predicting off-target effects in genetic modifications.

For supervised learning, I will simulate and create a synthetic dataset that mimics genomic sequences and off-target effects. This dataset will be inspired by existing genomic data to ensure that it resembles real-world data. I will research and incorporate known off-target sites and their sequences to train the model as effectively as possible. To introduce uncertainty and address diverse genomic sequences, I will introduce mutations, insertions, and deletions at random locations in the sequences. The synthetic dataset will be labeled using the off-target scores from the "CRISPRseek" database. Incorporating known off-target sites and their sequences from the database will help the model learn and predict off-target effects more accurately.

3.3 Implementation

To implement my AI algorithm, I will use PyTorch, a framework useful for creating and implementing deep learning models. First, I will use built-in libraries such as torch, and pandas to help load the "CRISPRseek" dataset, as well as create the synthetic dataset. Then, I will use the pandas library to apply one-hot encoding to the "targetSeqName" column to convert the column input into a more suitable format for this AI model. Next, I will instantiate the AI model as a neural network model, which is key in helping recognize patterns in the data and forming predictions it learned from the data. The model will be trained on both the synthetic dataset and the "CRISPRseek" dataset.

To store and analyze the results generated by the AI model, after each prediction, I will store the model's outputs, genomic sequences, predicted off-target effects, and other relevant information into a CSV file. By storing these results, I can further analyze the performance of the model, which will help determine if the model needs any refining. With the neural network model instantiated and trained on both datasets, the anticipated results will demonstrate a higher accuracy rate in predicting potential off-target effects.

4. ANTICIPATED RESULTS

Based on the design and implementation of the AI model, I anticipate promising results in predicting potential off-target effects in genetic modifications. Early-stage testing may indicate the AI model's ability to identify a significant number of off-target sites, making it a potential competitive tool against already existing algorithms such as CFD score, MIT score, and CROP-IT score.

After completing a comparative analysis on these algorithms, I anticipate the AI will demonstrate competitive performance, offering genetic engineers another tool in determining off-target predictions. This analysis will highlight the AI model's strengths and weaknesses compared to the other algorithms, aiming to be used as an alternative approach for predicting potential off-target sites with improved accuracy and reliability.

5. CONCLUSION

Genetic Engineering is a constantly evolving field of study. This project addresses the concerns for off-target effects in genetic modifications by proposing an AI-powered algorithm. This innovation offers enhanced accuracy and efficiency compared to models that already exist. By integrating advanced machine learning algorithms, including both unsupervised and supervised learning, the model aims to provide genetic engineers with a preliminary screening tool to assist them in making informed decisions about genetic modifications.

6. FUTURE WORK

Although the preliminary results received from the algorithm are promising, with the AI

model showing competitive performance compared to existing models, it is essential to conduct extensive testing and validation across a broader range of genetic contexts to ensure its reliability and accuracy. Collaborating with genetic experts and bioinformatics specialists will be crucial in refining and optimizing the model. Additionally, integrating real-time learning capabilities and continuously updating the model with new genomic data will be explored to enhance its predictive accuracy and adaptability. Once validated, the AI model has the potential to significantly enhance the safety and reliability of genetic engineering practices, ensuring more predictable and controlled outcomes.

REFERENCES

Bao, X., Pan, Y., Lee, C., Davis, T., & Bao, G. (2020). Tools for experimental and computational analyses of off-target editing by programmable nucleases. *Nature Protocols.* 16, 10-26

GPP—The Genetic Perturbation Platform. (2016). How the sgRNA Designer Works. https://portals.broadinstitute.org/gpp/public/software/sgrna-scoring-help#:~:text=The%20Cutting%20Frequency%20Determination%20(CFD,within%20the%20guide%20RNA%20sequence.

Lin, J., & Wong, K. (2018). Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics*, 34(17), i656-i663.

Singh, R., Kuscu, C., Quinlan, A., Qi, Y., & Adli, Mn (2021). Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Research*, 43(18), e118.

Zhu, L., Holmes, B., Aronin, N., & Brodsky, M. (2016). CRISPRseek: A Bioconductor Package to Identify Target-Specific Guide RNAs for CRISPR-Cas9 Genome-Editing Systems. Figshare. https://dio.org/10.1371/journal.pone.0108424.