Dean&Francis
ISSN 2959-409X

Applications and Future Perspectives of Artificial Intelligence in gRNA Design

Xiubin Zhang[1,*]

[1]Department of School of English and International Studies, Beijing Foreign Studies University, Beijing, 100089, China
*Corresponding author. Email: 202420101108@bfsu.edu.cn

Abstract:

The rapid development of CRISPR gene editing has brought revolutionary breakthroughs to the biomedical area, yet the design of gRNA is still facing problems as unstable editing efficiency and the risk of off-target. The introduction of Artificial Intelligence (AI) technology has guided this issue to a brighter future. Studies have shown that deep learning models can predict editing efficiency and assess off-target risk precisely through analyzing information as the sequence features of gRNA, while the introduction of tools like Reinforcement Learning (RL) further enlarges the space of gRNA design. There is still a deficiency in recent studies. This paper analyzes the application of AI in gRNA design, including editing efficiency and off-target risk assessment. The workflow, performance, and limitations of each model are concluded, and the conclusion of AI having a great devotion to gRNA design and the tips for choosing different models in different cases are given. This research provides references about choosing and designing models, but there are still unsolved challenges in aspects of building standard datasets and public databases. In the future, gRNA design can be optimized by reinforcing the use of Federated Learning (FL) and single-cell CRISPR screening technology.

Keywords: CRISPR; AI model; guideRNA design.

1. Introduction

CRISPR (Clustered Regularly Interspaced Palindromic Repeats) is a bacterial adaptive immune system, and Cas9 (CRISPR-associated protein 9) is an RNA-guided DNA endonuclease. CRISPR-Cas9 system helps people achieve the goal of gene editing through designing gRNA, delivery systems, making Double-Strand Break (DSB), and cellular repair, among which designing gRNA to an extent decides the specificity and efficiency of editing [1]. However, traditional gRNA design methods usually lead to a long timeline and high cost due to their high dependency on trial and error and empirical methods, and have difficulty meeting needs in a complicated genome environment.

So far, there are at least two bottlenecks in gRNA design. One is the unstable editing efficiency, that the same gRNA may have great activity variation across cell types or targets, sometimes it even goes up to 10 times than in the other. The other one is off-target

risk, which means the non-specific cleavage may lead to genome instability [2]. In 2014, the Doench lab for the first time used a logistic regression predictive model to recognize the sequence features of the top quintile gRNA selection by activity in each gene [1], and quantify the contribution weight gRNA sequence features have on editing efficiency through statistical methods. In 2016, this lab again designed Gradient Boosted Decision Trees (GBDT), an ensemble learning method that sequentially combines weak decision trees, each correcting the errors of the previous one, to a highly accurate one, and found that GBDT is the best among the tested models [3]. In 2018, the Chuai lab used a Deep Neural Network (DNN), a multi-layer artificial neural network that applies nonlinear activation functions at each step to extract increasingly complex and abstract features, and adjusted learning model via existing marked sgRNA and developed DeepCRISPR, an unsupervised pretraining model that optimizes sgRNA design in coding and non-coding regions [4]. In 2019, the Kim lab developed DeepSpCas9, a regression model based on deep learning and large-scale datasets. It predicts precisely the activity of SpCas9, which has high-generalization-performance [5]. And in 2025, the Qu lab developed proxy system called CRISPR-GPT that could understand and generate natural languages and structure text through deep learning as a Large Language Model (LLM), which could be used for automating and enhancing the gene editing design and data analysis based on CRISPR [6], AI technology has provided data-driven approach to gRNA design, and through finding rules behind a large scale of data, the aim of designing precisely is achieved.

This research will systematically elaborate on the development of AI optimizing gRNA design, divided into two parts: On-Target Efficiency Prediction and Off-target effects. By showing the workflow of each model and comparing them about their methods and performance, this paper will give suggestions on choosing a model in different cases and discuss challenges unsolved. In the end, the prospect would be given. This research aims to provide possible ideas about choosing tools for gene editing researchers, while at the same time promoting the development of related fields like precision medicine.

2. Efficiency Prediction

2.1 Concept

On-Target Efficiency Prediction is a computational or experimental estimation of how effectively a biomolecule (e.g., sgRNA, siRNA) induces edits in a particular mission. The key factors to efficiency are GC content, PAM-proximal bases, secondary structure, and epigenetic context. Balanced GC content (in 40%-60% [7]) is good for gRNA to combine with target DNA. It may cause the two to combine too tightly and lead to off-target effects or form a hairpin structure and hinder the combination if the GC content is too high. Contrarily, if it's too low, the binding affinity may be lowered and lead to low editing efficiency. Different Cas variants have different binding affinity to PAM-proximal bases, and the difference in bases' preference will also lead to variant cleavage efficiency. Thus, a strong binding context is recommended. Structures like hairpin can cause physical barriers to the binding of gRNA-Cas complex, and the stem-loop structure of gRNA can also affect the identification and binding of it and the Cas protein. Therefore, it's necessary to avoid stable secondary structures. A DNA sequence having too low chromatin accessibility may hinder the binding of the Cas protein; thus, chromatin open regions are recommended. Assay for Transposase- Accessible Chromatin Sequencing (ATAC-seq) data can help design gRNA [8]. Table 1 shows the mechanism and the optimization strategy of each factor.

Table 1. Mechanism and the optimization strategy of each factor

Factor	Mechanism	Optimization strategy
GC Content	-detect GC content (40%-60% is recommended) [7]	Choose medium content GC
PAM-Proximal Bases	4-6 bases in NGG downstream can affect Cas9 binding [6]	A strong binding context is recommended
Secondary Structure	Self-folding of sgRNA may hinder Cas9 binding [6]	Predict and avoid stable secondary structure
Epigenetic Context	Chromatin open regions are more likely to be edited [8]	Choose a target with the help of ATAC-seq data

ISSN 2959-409X

2.2 Efficiency Prediction Model

2.2.1 From Traditional Methods to Deep Learning

Early rule-based models like CRISPR predict efficiency by concluding empirical rules based on sequence features of high-efficiency gRNAs (e.g., GC content, PAM-proximal bases, secondary structure, epigenetic context, and 20 more features in total). The benefits of using this method are fast calculation and the need for no training data. But it only suits small-scale screening while neglecting the variance of epigenetics and cell types. Later, the development of linear regression models made efficient prediction more precise than using rule models, but it still depends on the training data of particular experiments and lacks cross-species generalization. Models that incorporate epigenetic data, like chromatin accessibility and histone marks, into prediction are called Epigenetic-aware Models. DeepCpf1, developed by the Kim lab in 2018, improves the prediction accuracy of complex cell types significantly, but its computational complexity is high, and the cost to users of providing data is also too high [9]. Now, deep learning models are becoming a trend. DNN models like DeepCRISPR and Bidirectional Recurrent Neural Network (BiRNN) models like PRIDICT are also popular. Here, this research will take PRIDICT as an example, introduce its workflow, performance, and compare it with other models to better make readers understand efficiency prediction models [10].

2.2.2 PRIDICT

PRIDICT (PRIme editing guide preDICTION) is a BiRNN model based on Attention (AttnBiRNN). Here is its workflow.

First are data collection and preprocessing. It is divided into two parts: High-Throughput Screening (HTS) experiments and feature extraction. The large-scale editing efficiency of pegRNA in human pathogenic mutations is analyzed through a self-targeting library, and every pegRNA or target site that has more than 100 reads is retained. Sequences that have high editing rates are excluded, and the final dataset is ready. Then extract 67 features from pegRNA and target sequence, including sequence features (e.g., GC content, poly-T length, RTT length, PBS Tm, and edit type) and structural features like Minimum Free Energy (MFE) and bases near the edit site.

Next is the formal workflow of design PRIDICT. It includes two plates: one is the Encoder, and the other is the Decoder. Encoder is composed of two sequence encoders and one feature encoder. The sequence encoders process the raw sequence and the edited sequence, respectively, by using Bidirectional Gated Recurrent Units (BiGRUs), and then weight important sequences dynamically through Attention. The feature encoder processes 67 predefined features through a free-forward network. The Decoder is responsible for mapping the output of the three encoders into three types of probability distributions: intended edit, unintended edit, and no edit.

Then there is training and validation. 5-Fold Cross-Validation is applied to group data by treatment site into a training set (80%) and a test set (20%), in which the training set is secondly grouped and 10% of it is token out randomly as a validation set. The core of 5-Fold Cross-Validation is dividing the dataset into 5 pieces, and choosing one of them as the test set, the others as training sets in turn, and getting the average results after 5 times repetition. Using this validation method can avoid pegRNA from polluting training sets and test sets effectively. Training sets are used for model training, while the test set is used for evaluating model performance independently, and the validation set is used for hyperparameter optimization.

Next is comparing PRIDICT with several trained traditional machine learning models, such as linear regression model Lasso, Ridge, and Elastic Net, Random Forest, XGBoost, and Histogram-Based Boosting. The result shows that among the traditional machine learning models, XGBoost performs the best (Spearman's R=0.80), while PRIDICT further improved performance (R=0.85).

In the end is the validation and application part. Endogenous locus validation shows that the editing efficiency of 45 tested pegRNA in HEK293T and K562 cells, getting a score over 70, is 10-20 times better than before, and external data validation also gets good feedback (in the data of Anzalone et al. and Kim et al., PRIDICT performs better than other tools). Adaptive validation also gets good results as PRIDICT shows better performance in MMR-deficient cells. It's the same in vivo testing. In mouse hepatocytes, the pegRNA editing efficiency, having score over 70 of PRIDICT is raised by 5.9-9.6 times.

3. Off-Target Risk Assessment

3.1 Off-Target Effects

Off-target effects refer to unintended DNA cleavage at genomic sites with partial homology to the gRNA target sequence, potentially causing gene disruption, chromosomal translocations and epigenetic dysregulation, and other unpredictable outcomes. There are many reasons that can cause off-target effects, including improper design of gRNA (like false matching with DNA and causing partial complementarity to non-target sequences), dependence and compatibility of editing tools like CRISPR-Cas9 to PAM sequence, and the difference in cell types causes dif-

ferent chromatin open states or DNA repair mechanisms that affect editing specificity. Target site prediction can be searched via sequence similarity, or filtered to exclude tight chromatin regions through chromatin accessibility and remain potential off-target sites in open regions, or by conducting energy model evaluation, such as calculating gRNA-DNA hybrid free energy [4,9]. Workflows of most off-target risk assessment models are the same in general, including inputting the characteristics (e.g., mismatch number and position, local GC content, secondary structure prediction, and epigenetic features), off-target site identification, feature integration, and risk probability calculation. Different models may have variance in details, but the overall trend is consistent. Here, this research will take three models, DeepCRISPR, CHOPCHOP v2, and Cas-OFFinder as examples, introduce their workflows, performances, and applicable scenarios respectively, and by comparation [11,12].

3.2 Off-Target Risk Assessment Prediction Models

3.2.1 Cas-OFFinder

The core aim of Cas-OFFinder is to identify potential off-target sites in the genome rapidly. Its workflow can be divided into three steps: input, parallelized search, and output.

First of all, the user needs to provide sgRNA sequence and PAM (e.g., NGG of SpCas9) and set the maximum mismatch tolerance to start parallelized search. Based on the open parallel programming frameworks OpenCL (Open Computing Language), Cas-OFFinder assigns computing missions to several hardware accelerators to conduct them simultaneously and perform parallel chunked alignment of the genome. It supports CPU/GPU acceleration and is able to handle large-scale genomic data; at the same time, Cas-OFFinder's flexible PAM matching mechanism allows it to be compliant with various Cas9 variants. After binding sgRNA and engineered Cas9 variants into potential 23-bp target sequences, genome data are loaded in chunks, and matching candidate sequences are searched in parallel. Then 20bp region in sgRNA is compared to record mismatch numbers, and in the end, the off-target sites (within the mismatch threshold set by the user) are outputted, and the position, direction, and mismatch of chromosomes are marked.

3.2.2 CHOPCHOP v2

The core aim of CHOPCHOP v2 is to provide user-friendly sgRNA designs that are available to support multiple CRISPR systems. The workflow of it could be divided into four steps: input, target screening and scoring, off-target detection, and output.

First, the user should input gene ID, coordinates, or sequences, and select the target region. Then the GC content will be screened to avoid coiled A/T, and use Dual Nickase mode to select sgRNA pairs spaced 10-31bp apart and truncate them into 17-20bp length to lower off-target, balanced target range and specificity. Next, conduct off-target prediction and filter low-risk sites. In the end, the sgRNA quality metrics will be output by an interactive plot, and the information on validation of primers and restriction enzyme sites will be generated automatically.

3.2.3 DeepCRISPR

The core aim of DeepCRISPR is to predict the knockout efficiency of sgRNA and genome-wide off-target profiling through deep learning. This model can be divided into on-target efficiency and off-target effect prediction; here, only the second part is discussed.

A Dual-Channel Hybrid Deep Neural Network is applied in DeepCRISPR, which is carried out in two stages: an unsupervised pretraining stage and a supervised fine-tuning stage, and each stage is carried out in three steps: input, model processing, and output. First, prepare and preprocess data. Input 20 bp target sequence and NGG PAM sequence, and pre-screening sites that have no more than 6 mismatches with sgRNA in the whole genome. Then use 4-channel one-hot encoding to denote sgRNA and off-target site sequences, like A is [1,0,0,0], C is [0,1,0,0], and so on. And then enhance the off-target data that have been validated by several tests, and expand the training sets by introducing validated off-target sites with random mismatches.

Unsupervised pretraining stage uses Deep Convolutional Denoising Neural Network (DCDNN), namely using a Multi-Layer Convolutional Neural Network (CNN) structure to automatically extract high-dimensional representation of sgRNA sequence and epigenetic features, and add random noises into the data inputted during the training process to force the model to learn Robust feature representations to avoid overfitting. In the end, the input data are compressed into low-dimensional features by the encoder for later fine-tuning. About 0.68 billion unmarked sgRNA sequences are training DCDNN in this stage to let the model learn common sequence and epigenetic feature patterns.

3.3 Model Performance Evaluation and Comparative Analysis

Considering that each model uses a different method above, the applicable scenarios for each model are varied. Cas-OFFinder has a rapid operating speed, while it only provides potential off-target sites and has no activity pre-

ISSN 2959-409X

diction, so it's suitable for initial screening of off-target sites. CHOPCHOP v2 depends on manual rules; it has limited prediction accuracy. But since it has strong flexibility and supports various CRISPR tools, it's suitable for the whole process design of the experiment. DeepCRISPR is an automated feature learning. It has strong cell generalization capability and is applicable for high-precision sgRNA optimization and cross-cell-type prediction. Table 2 is the outcome of the comparison of the three models.

Table 2. The outcome of the comparison of the three models

Model	Method	Strength	Limitation
DeepCRISPR [10]	CNN	High-Precision Prediction, Automated Feature Learning, and Cross-Cell-Line Generalization	High Data Dependency and Low Small-Sample Performance
CHOPCHOP [11]	Rule-Based+ Machine-Driven Hybrid Approach	Fast and Interpretable	Incomplete Off-Target Prediction
Cas-OFFinder [12]	Brute-Force Search	Fast and Flexible	DNA Sequence Only and a large amount of computation

4. Conclusion

This research systematically concludes the development of AI research in gRNA design and focuses on analyzing the application of AI in fields like efficiency prediction and off-target risk assessment. The workflow, performance, strengths, and limitations of each model are shown, and suggestions about choosing a different model in an applicable scenario are given. The result shows that strategies based on BiRNN and RL are able to approach needs that aim to realize an effectively balanced multi-objective design and offer a reference to studies that focus on gRNA design in complex environments in the future.

The core meaning of this research is to reveal the large potential of AI technology in the gene editing field. The data-driven paradigms of AI models, in some ways, fill the deficiency of traditional rule-based methods and provide a reference to gene editing for personalized medicine and precision agriculture. Moreover, the off-target effect prediction AI models also give new ideas to lower gene therapy risks. However, studies by now still have some limitations, such as the models still depending on high-quality training datasets, which, to some extent, limit their application range.

In the future, researchers can focus on the application of transfer learning in cross-species gRNA design and develop more explainable AI methods like Attention, to help deepen understanding of gRNA functional mechanisms, and lay the foundation for the transfer of gene editing technology to the clinic and industrial applications.

References

[1] Doench J G, et al. Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nature Biotechnology, 2014, 32: 1262-1267.

[2] Fu Y F, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nature Biotechnology, 2013, 31: 822-826.

[3] Doench J G, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature Biotechnology, 2016, 34: 184-191.

[4] Chuai G H, et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biology, 2018, 19: 80.

[5] Kim H K, et al. SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. Science Advances, 2019, 5: eaax9249.

[6] Qu Y H, et al. CRISPR-GPT for agentic automation of gene-editing experiments. Nature Biomedical Engineering, 2025.

[7] Hsu P D, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nature Biotechnology, 2013, 31: 827-832.

[8] Ranzoni A M, et al. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. Cell Stem Cell, 2021, 28: 472-487.

[9] Kim H K, et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nature Biotechnology, 2018, 36: 239-241.

[10] Mathis N, et al. Predicting prime editing efficiency and product purity by deep learning. Nature Biotechnology, 2023, 41: 1151-1159.

[11] Labun K, et al. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. Nucleic Acids Research, 2016, 44: W272-W276.

[12] Bae S, Park J, Kim J S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics, 2014, 30: 1473-1475.