Proceedings of the 3rd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/2024.KA20158

Machine Learning-Based Risk Assessment of Gene Editing: Degeneration During Gene Editing Possibility of Target Effect

Jingbo Wang[1,a,*]

[1]*Department of Engineering, Shenzhen MSU-BIT University, Shenzhen, 518000, China*
a. 1120210420@smbu.edu.cn
**corresponding author*

Abstract: CRISPI-Cas 9 technology is currently the most popular gene editing tool, which has been used in biomedical research and clinical treatment, but it also has the problem of off-target effects that cause unexpected gene mutations, which limits its safety and efficacy. Now, various machine learning models have been trained to predict off-target events in order to: Gene editing is more precise. In terms of both accuracy and interpretability there are still issues in the existing models. This paper has focused on reviewing Machine learning archery Off Target Prediction models and their performance. The results of this study show that the deep learning model has high accuracy and interpretability in the prediction of off-target events, which provides a reference for the optimization of gene editing strategies. The results reveal that there is a lot of room for development of deep learning technology in improving the safety of gene editing, which provides an in-depth reference for the sustainable development of gene editing technology. The application of this study to this paper also has its limitations, including the inadequacy of the applicability of the model in multiple gene editing contexts. In the future research of gene editing, more attention can be paid to the study of new biological indicators, the establishment of more accurate prediction models, and the improvement of models with more explanatory interpretation, so as to more safely apply gene editing technology in clinical and agricultural treatment and improvement.

Keywords: CRISPI-Cas 9, off-target effects, machine learning, deep learning.

1. Introduction

The CRISPR-Cas9 system, born in 2012, has become a scientific breakthrough in the field of life sciences. It can be efficiently edited at this allele by selecting a simple sgRNA, and the CRISPR/Cas9 system has higher editing efficiency and higher target specificity than the previously commonly used zinc finger nuclease (ZFN) and activator of transcription effector nuclease (TALEN) methods [1]. This technology has shown great promise in biomedical research, agricultural biotechnology, and potential clinical treatments, including genetic disease treatment, cancer research, personalized medicine, and more.

However, in the consideration of CRISPR-Cas9 technology, off-target effects inevitably exist, that is, gene editing is not the target location, resulting in mutations at non-gene sites, which not only affects the target nature and specificity of gene editing, but also poses safety risks. The existence of

DOI: 10.54254/2753-8818/2024.KA20158

off-target effects has greatly limited the further expansion and clinical application of the CRISPR/Cas9 system [2]. Because it may cause chromosomal rearrangements and cause damage on inexperfectly matched gene bodies, it makes the application of gene editing technology in human medical research more difficult. To address this issue, a number of methods have been found to precisely control the CRISPR genome editor, and strategies have been developed to reduce off-target effects, including optimizing sgRNA design, developing high-fidelity Cas9 variants, and applying Cas9 orthologs [3]. A large number of gene editing datasets are obtained through computer science algorithms, and models can be used to predict off-target events by using machine learning models, and risk assessment and gene editing strategies can be optimized for researchers.

The off-target effects of gene editing have been studied for a long time, and the significance is to facilitate a better understanding of the off-target mechanism of the CRISPR-Cas9 system, and then develop safer and more accurate gene editing tools to promote the application of gene editing technology in clinical treatment, agricultural improvement and other fields. With the advancement of artificial intelligence technology, the accuracy and efficiency of gene editing will be improved even higher, and it will be brought from the laboratory to the clinic, with the potential not only to create new breakthroughs and change the fate of human medicine, but also to improve agricultural production and cope with climate and ecological challenges.

2. The possibility of off-target effects

CRISPR-Cas9 is a new gene-editing technology that works by genetically modifying the genome using bacterial immune defense mechanisms. The system consists of a single guide RNA (SgRNA) and a Cas9 nuclease, which directs the Cas9 enzyme to a specific DNA target site, binds the target DNA sequence through adapter bases, and then the Cas9 enzyme cleaves the DNA duplex at the target location under the command of the sgRNA [4,5]. CRISPR-Cas9 is a precision DNA cutting technology that makes it a key tool for studying gene function, gene therapy, and agricultural biotechnology.

Off-target effects are a common problem with interventions using the CRISPR-Cas9 system, and they cannot be avoided issue. An off-target effect is a type of SgRNA-induced DNA cleavage that occurs at a non-target location, resulting in inappropriate forms of gene mutation and gene expression inactivation, which hinders the cells from performing their normal functions and can sometimes even lead to malignant damage such as carcinogenesis. Regarding the biological mechanism of off-target effects, the main thing is the differential pairing between the SgRNA and the target DNA, and within the allowable range of matching between the sgRNA and the target DNA, the Cas9 enzyme may still cleave elsewhere because of the mismatch of several nucleotides. The off-target effect is also involved with the PAM sequence recognition profile since PAM sequence (Protospacer Adjacent Motif) is a prerequisite for Cas9 enzyme to bind DNA [6]. The possibility of off-target effects can also be affected by the design of sgRNAs, the specificity of Cas9 enzymes, and the mechanism of DNA repair in cells. Other DNA repair pathways, such as non-homologous end joining (NHEJ) and micro-homologous end joining (MMEJ), can also interfere with CRISPR-Cas9-induced DNA fragmentation and thus affect the occurrence of off-target effects [7]. From the above analysis, it can be seen that understanding and controlling for off-target effects is an extremely critical issue to ensure the safety and efficacy of the CRISPR-Cas9 system.

3. The application of machine learning in off-target prediction

Prediction of off-target effects in psychiatric disorders is critical to the modern drug discovery process, as off-target effects determine the safety and efficacy of drugs. A new machine learning framework, FRoGS (Functional Representation of Gene Signatures), which uses deep learning to predict the

DOI: 10.54254/2753-8818/2024.KA20158

efficacy target of a drug. Genetic signatures translate their names into their biological functions, similar to the word2vec technique in natural language processing. This method not only uses the well-known gene function in Gene Ontology (GO), but also comes with experimental expression data from the ARCHS4 database for training for deep learning models. In this way, FROGS can transform genes into a high-dimensional space and bring genes with similar functions close together in the vector space, providing a better way for them to capture genes with similar functions. For the signature vectors that compare compound perturbations and target gene regulation, the similarity between them can be calculated directly, so the Siamese neural network model will be used. With such balanced training approach, the model can generalize well between known and unknown targets [8].

DIPOFF model integrated deep learning and explainability and can predict off. target sites of CRlSPR-Cas9along with the help of deep learning [1]. This model uses recurrent neural networks (RNNs), in which long short-term memory (LSTM) and gated recurrent units (GRU) can better process sequence data and capture long-term dependencies in sequence data. In the data preprocessing section, the researchers used the One-hot Encoding contrast coding method to preprocess the sgRNA and target DNA sequences, and added a new orientation channel to represent the parts that did not match the sequence. These coding schemes allow the model to learn about the relationship between the SgRNA and the target DNA. The main purpose of this is to convert the input sequence into data in a numerical format that the deep learning model is capable of processing. Then it goes to the genetic algorithm to improve the hyperparameters of the model for more accurate off-target predictions. This model also used a general parametric RNN model which consists of one or two Recurrent layers, where each layer may be unidirectional or bidirectional. The output of the model is in a layered structure, and one layer can be seen as the result of connecting to the next layer and ultimately to the output layer. During this process, cross entropy was employed as the form and the former one was trained with Adam optimizer. The batch normalization and the dropout lavers were incorporated into this model in order to reduce overfitting. By way of such the explanation, the study revealed that there were two subregions within SgRNA seed region which is associated with off-target effects providing new biological insights into the off-target mechanisms of CRISPR-Cas9 [9].

In transforming the morphology of sgRNA-DNA pairs, the feature sequence isautomatically trained using the deep-learning model of CnnCrispr: the word vectors are trained with the GloVe model [10], and then integrated into the deep model comprising biLSTM and the 5-laver CNN .The innovation of the CnnCrispr model lies in the fact that it can reveal patterns of interaction between sgRNA and target DNA. Through the GloVe model, CnnCrispr can convert sequence information into vector representations that can reflect the degree of association of a particular sequence with other sequences, what biologist calls inter-sequence regularization information.Besides,ability to capture sequence context information in the biLSTM layer is critical for understanding the complex interaction patterns between sgRNA and DNA in the proposed model. In the computer science level, the CnnCrispr model uses several technologies from the depths of learning to: biLSTM networks process sequence data to capture long-term dependencies in sequences; CNN extracts local features through multi-layer convolutional layers, and synthesizes global information at the high level. In order to prevent earthquakes, Batch Normalization and Dropout layers are also introduced into the model to improve the generalization ability of the model. By virtue of this structure design, CnnCrispr is capable of handling highly imbalanced datasets while maintaining high prediction accuracy, making it possible for CnnCrispr to learn the sequence features of sgRNA and potential off-target sites, as well as automatically predict the off-target tendency of a particular DNA segment, it has also avoided the noise or interference information that may be brought in the process of manually constructing features [11].

The ABE deep off and CBE deep off models, capable of predicting potential off-target sites, employ a fusion embedding approach, treating the gRNA and off-target sequences as the sole

DOI: 10.54254/2753-8818/2024.KA20158

sequence inputs and encoding them in a real-valued matrix space. This design allows the model to improve generalization and training speed by sharing weight initialization settings.an attention mechanism is applied to the model to derive more representative features from LSTM (Long Short-Term Memory Network) outputs and input these features into a fully connected neural network. As these models can handle a vast amount of sequence data and identify key sequence features affecting off-target efficiency, they enable the understanding of base editor specificity by comparing the impact of different mutation types, such as mismatches, insertions, and deletions, on off-target efficiency, thus providing guidance for experimental design. In terms of post-processing interpretability of the models, the "Layer Integrated Gradient" class is utilized to assess the attribution scores of each nucleotide position in the input sequence. These estimates indicate that traits in mutation sites have a negative impact on non-target ratios, while traits in matching sites have a lesser effect [12].

4. Evaluation

4.1. Limitations

Although previous studies have demonstrated that the application of deep learning can enhance the ability to predict off-target effects, they often struggle with the trade-off between precision and completeness, which limits their effectiveness and also fails to provide correct explanations for the complex decision-making processes of their models [13].

4.2. Comparative Analysis

CRISPR-DIPOFF was compared with other non-target prediction models such as CNN_Std [14], Deep CRISPR [15], AttnTo Mismtach_CNN [16], CnnCrispr [11], CRISPR-IP, FNN_8Ch, and ELECTRA [17] to demonstrate the model's performance in the trade-off between accuracy and memory completeness (compromise) to achieve a balance of efficiency and explainability. However, in the base model, CnnCrispr had the most balanced performance, followed by CRISPR-IP. Among them, the LSTM and GRU CRISPR-DIPOFF models outperformed all other models in terms of AUPRC (Area Under the Precision-Remember Curve), showing the best performance. In particular, the LSTM model showed a significant improvement in F1 scores and AUPRC scores compared to the CnnCrispr baseline model. These results indicate that sequence-based models and, hyperparameter-optimized models enhanced the ability to predict non-target CRISPR-Cas9 values [13].

CnnCrispr demonstrated superior predictive performance across multiple test sets, with higher recall values compared to other pre-selected models. CnnCrispr is the first model to use only sequence information to predict non-target propensity, thus avoiding incorrect or confusing information that can be introduced by artificially generated features, making prediction results more reliable. In the “leave one sgRNA” experiment, the auROC and auPRC indices acquired by CnnCrispr were higher on average than those of the other four models: CFD [18], MIT [19], CNN_Std [14] and DeepCRISPR [15] we can also plot its accuracy. as well as generalizing ability in estimating off-target propensities of the novel sgRNAs [11].

5. Experimental Validation

Regarding models selection, the experiment analyzed many kinds of Recurrent Neural Networks and its derivatives, traditional RNNs, LSTM (Long Short Term Memory), GRU (Gated Recurrent units), and bidirectional and stacked, and so on. The hyperparameters of the model are optimized by genetic algorithms, which include the parameters settings, model training, fitness estimation, model selection and then cross and variation of parameters to generate one set of parameters [13].

DOI: 10.54254/2753-8818/2024.KA20158

In terms of performance validation, CnnCrispr underwent a "leave-one-sgRNA-out" cross-validation on the Deep Crispr dataset, and the results showed an auROC (area under the receiver operating characteristic curve) of 0.957 and an auPRC (area under the precision-recall curve) of 0.429 in the classification model, demonstrating superior classification and regression performance compared to existing state-of-the-art models [11].

In terms of model development, gRNA-non-target pairs targeting adenine base editors (ABEs) and cytosine base editors (CBEs) have been developed and stably integrated into human cells. After five days of editing, a valid off-target dataset of 54,663 ABE and 55,727 CBE was obtained, which were used to train the deep learning model, and finally two models, ABE deep off and CBE deep off, were developed [12].

In terms of data selection for application, to validate the predictive power of the FRoGS model, researchers applied the model to predict all compound-gene pairs, not just those with known target annotations in the L1000 database. With this method, 780,438 complex target pairs were predicted, and the probability values of these pairs were above 0.8. These predictions have been validated in a variety of ways, including using historical pIC50 data from Novartis, probabilistic structure-activity relationship analysis (SAR), and activity data from the public NCI60 and PSP databases [8].

6. Conclusion

This article conducts an in-depth analysis of the off-target effects in CRISPR-Cas9 gene-editing technology, explores the application of machine learning models in off-target prediction, and compares the performance of various models. The research indicates that machine learning, particularly deep learning, demonstrates significant potential in predicting the Possibility of off-target events and in optimizing gene-editing strategies. The insights from this article are expected to promote the safe application of gene-editing technology in fields such as clinical therapy and agricultural improvement, while also providing new perspectives and methodological references for future research. Despite some progress in predicting and modelling non-target effects, there are still limitations. For instance, the generalizing ability of models and their explanatory ability to explain complex biological mechanisms need to be further developed; there is a lack of analysis on the adaptability and robustness of models in different gene-editing scenarios. With the enhancement of computational power and innovation in algorithms, it is anticipated that the application of machine learning models in the field of gene editing will become more extensive. Future research can focus on developing more accurate predictive models, exploring new biological features, and improving model interpretability. Interdisciplinary collaboration will help integrate knowledge from different fields, thereby providing more comprehensive safeguards for the safety and efficacy of gene-editing technology.

References

[1] *Wood A.J.,Lo T.W.,Zeitler B.,Pickle C.S.,Ralston E.J.,Lee A.H.,Amora R.,Miller J.C.,Leung E.,Meng X.,et al.(2011).Targeted Genome Editing Across Species Using ZFNs and TALENs.Science,333,307.*

[2] *ZHANG Chen, LEI Zhan, LI Kai, SHANG Ying, XU Wen-tao.(2020).Research Progress on Off-target Effects and Detection Techniques in CRISPR/Cas9 Systems[J]. Biotechnology Bulletin, 36(3): 78-87.*

[3] *Chen K., Han H., Zhao S., et al.(2024).Lung and liver editing by lipid nanoparticle delivery of a stable CRISPR–Cas9 ribonucleoprotein.Nat Biotechnol.*

[4] *Jinek, M. et al. (2012).A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity . Science Magazine,Vol 337,Issue 6096,pp.816-821.*

[5] *Hsu P.D.,et al.(2013).DNA targeting specificity of RNA-guided Cas9 nucleases.Nat Biotechnol.*

[6] *Chen J.S., et al. (2017).Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. Nature.*

[7] *Kim D., et al. (2015).Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nature Methods.*

DOI: 10.54254/2753-8818/2024.KA20158

[8] Hao Chen,Frederick J.King,Bin Zhou,Yu Wang,Carter J.Candy,Joel Hayashi,Yang Zhong,Max W.Chang.et al.(2024).Drug target prediction through deep learning functional representation of gene signatures.nature communications,15:1853.

[9] Md. Toufikuzzaman , Md. Abul Hassan Samee and M. Sohel Rahman.(2024).CRISPR-DIPOFF: an interpretable deep learning approach for CRISPR Cas-9 off-target prediction.Briefings in Bioinformatics, 25(2), 1–10.

[10] Pennington J.,Socher R.,Manning CD.(2014)GloVe: Global Vectors for Word Representation.Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),p.1532–43.

[11] Liu Q., Cheng X., Liu G., et al.(2020) Deep learning improves the ability of sgRNA off-target propensity prediction. BMC Bioinformatics,21(1):1–15.

[12] Chengdong Zhang, Yuan Yang, Tao Qi , Yuening Zhang , Linghui Hou , Jingjing Wei , Jingcheng Yang , Leming Shi, et al.(2023)Prediction of base editor off-targets by deep learning.Nature Communications,14:5358.

[13] Toufikuzzaman Md. , Abul Hassan Samee Md. and Sohel Rahman M.(2024).CRISPR-DIPOFF: an interpretable deep learning approach for CRISPR Cas-9 off-target prediction.Briefings in Bioinformatics,25(2), 1–10.

[14] Jiecong Lin and Ka-Chun Wong.(2018).Off-target predictions in CRISPR-Cas9 gene editing using deep learning.Bioinformatics, 34, 2018, i656–i663.

[15] Chuai G., Ma H., Yan J., et al.(2018) Deepcrispr: optimized crispr guide rna design by deep learning. Genome Biol,19:1–18.

[16] Liu Q., He D., Xie L.(2019) Prediction of off-target specificity and cellspecific fitness of crispr-cas system using attention boosted deep learning and network-based gene feature. PLoS Comput Biol,15(10):e1007480.

[17] Clark K., Luong M-T., Le QV., Manning CD.(2020)Electra: pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:200310555.

[18] Doench JG., Fusi N., Sullender M., Hegde M., Vaimberg EW., Donovan KF., Smith I., Tothova Z., Wilen C., Orchard R. (2016).Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol,34(2):184–91.

[19] Hsu PD., Scott DA., Weinstein JA., F. Ann R, Silvana K., Vineeta A., Yinqing L., Fine EJ., Xuebing W., Ophir S.(2013).DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol,31(9):827–32.