Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/60/20241417

The integration of machine learning and CRISPR/Cas technology in the field of COVID-19 diagnostics

Yuezhe Qi

School of Agriculture, Huazhong Agricultural University, Wuhan, China

1811040409@stu.hrbust.edu.cn

Abstract. The coronavirus known as severe acute respiratory syndrome coronavirus 2(SARS-CoV-2), has generated a global health crisis that has attracted a lot of attention. The necessity of tracking and identifying these variants is highlighted by the appearance of several SARS-CoV-2 variations. Polymerase Chain Reaction-based(PCR-based) assays are considered the gold standard for virus detection, but they have many limitations. The Clustered Regularly Interspaced Short Palindromic Repeats(CRISPR-Cas) system has the potential to overcome these limitations, offering rapidity, sensitivity, specificity, and programmability. Machine learning is also a well-used technology that is widely employed in various fields. Both users and developers of the CRISPR/Cas toolkit have profited from the latest advances and adaptations of its techniques in the field of gene editing. This article reviews the application of CRISPR-Cas and machine learning in SARS-CoV-2 detection. The principles of the CRISPR-Cas system are elucidated, and the application of the combination of the two in detection is explored, providing insightful observations into the advancement of techniques for detecting mutations. In addition, their systematic applications, strengths and weaknesses, challenges and prospects are analyzed.

Keywords: CRISPR-Cas, Machine learning, SARS-CoV-2, diagnostics.

1. Introduction

In early December 2019, an outbreak of the respiratory disease Coronavirus Disease 2019 (Covid-19), caused by the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2), received considerable attention as a potential threat to global public health security. The necessity of continuing to prevent and control of the outbreak is still critical even if it was declared that the situation no longer qualifies as a Public Health Emergency of International Concern (PHEIC). The appearance of various SARS-CoV-2 variants highlights the importance of monitoring and recognising these variants. The creation of quick and trustworthy diagnostic methods for the novel coronavirus is an essential step in preventing further infections [1].

Among the available techniques, PCR-based assays are regarded as the gold standard for virus detection, offering high sensitivity and specificity. However, they are not without limitations, including the need for high-purity samples, expensive laboratory equipment, expert training, and a lengthy reaction time.

The discovery of CRISPR technology and related proteins (Cas) has considerably broadened the scope for effective molecular diagnosis and gene editing, thereby opening new avenues of enquiry in

DOI: 10.54254/2753-8818/60/20241417

these fields. CRISPR technology's programmability and adaptability enable it to target distinct gene sequences through sequence modifications in the CRISPR RNA (crRNA) sequence. Consequently, in response to the emergence of new mutations, the utilization of the CRISPR system enables expeditious adaptation and the design of crRNAs to aim for particular mutation sites, thereby facilitating the diagnosis of diverse SARS-CoV-2 variants in response to the prevalence of new variants [2].

In contrast, machine learning is the scientific discipline concerned with the use of computers to simulate or implement human learning activities, with a particular focus on the effective utilization of information, and the acquisition of hidden, valid, and comprehensible knowledge from vast quantities of data. Given the optimization capabilities of artificial intelligence and the surge in viral genomic data, optimization methods will facilitate the design of more sensitive virus detection than existing methods. Furthermore, these methods will enable the rapid design of new assays that provide new detection methods in viral mutations [3].

In the context of CRISPR-based detection of new coronaviruses, the creation of innovative machine learning models for prediction and analysis tools is of great importance. These tools can be applied to the design of good guide RNAs(gRNA)and the prediction of their activity. This paper provides an overview of systems for the detection of novel coronaviruses using CRISPR-Cas and offers insights into the use of machine learning algorithms for the detection of mutations.

2. CRISPR/Cas Technology System

The clustered regularly interspaced short palindromic repeats/CRISPR-associated proteins system, Prokaryotes' acquired immune system, is employed to resist the invasion of exogenous genetic components found in plasmids or phages. It is currently being widely utilized in genetic engineering, and also has significant applications in the diagnosis of new coronaviruses.

2.1. Mechanism of CRISPR/Cas System in gene editing and viral detection

2.1.1. Overview of the CRISPR/Cas system

Derived from the adaptive immune mechanisms of bacteria and archaea, the CRISPR-Cas system has rapidly evolved into an indispensable tool for biomedical research in biotechnology and for genome editing purposes. This revolutionary technology exploits the inherent capabilities of these systems, resulting in remarkable specificity and high efficiency for targeted modification of DNA sequences[4].

CRISPR and Cas are the two main parts of this system. Bacterial genomes have short repeating sequences called CRISPR sites, which are broken up by distinct spacers that are obtained from previous invaders like viruses and plasmids. A precursor CRISPR RNA (pre-crRNA) is generated by transcription of the locus and is subsequently processed into mature CRISPR RNAs (crRNAs). In order to provide adaptive protection against bacteria, these crRNAs operate as guides for Cas proteins, telling them where to locate and cleave complimentary sequences in invasive DNA.

2.1.2. Genome editing capabilities

CRISPR-Cas9 is a prime example of a type II CRISPR-Cas system that has attracted a lot of attention due to its versatility and ease of use for genome editing. In this system, a single guide RNA (sgRNA) is formed when a crRNA attaches to a trans-activated crRNA (tracrRNA). Following its direction by the sgRNA, the Cas9 nuclease cleaves the target DNA sequence, causing an insertion or deletion that disrupts the target gene. The process of gene editing can be completed through the subsequent repair or linkage.

2.1.3. Adaptation for viral detection

At present, CRISPR-based nucleic acid detection can be classified into two principal categories.

(1) Utilising Cas proteins, including Cas9 and other Cas proteins, to achieve highly specific recognition and binding of dsDNA (double-strand DNA);

DOI: 10.54254/2753-8818/60/20241417

(2) Leveraging the property that, subsequent to the specific recognition of nucleic acids by Cas proteins, For example, In Cas12 or Cas13 nucleic acid detection systems, upon binding of the crRNA and Cas protein complex to the target (Cas12a binds dsDNA, Cas13a binds RNA), the Cas protein bypass cleavage activity is activated, which in turn non-specifically cleaves the reporter probes (Cas12a cleaves ssDNA, Cas13a cleaves ssRNA). The reporter probes can be labelled with fluorescent groups, or with other quenching groups in order to convert the assay results into visual fluorescent signals or test paper strips.

2.2. CRISPR/Cas Diagnostic Techniques for COVID-19

2.2.1. SHERLOCK (Specific High-sensitivity Enzymatic Reporter UnLOCKing) system

SHERLOCK, represents a nucleic acid detection instrument developed by Feng. In contrast to DETECTR, which is based on the Cas12 system, SHERLOCK is a widely employed method for the detection of RNA viruses. This is because the activation of the Cas13 component results in the cleavage of the RNA molecule that is bound to the reporter molecule, thereby confirming the presence of the virus. The entire process of detection using the Cas13a detection platform is shown in Figure 1. Following the extraction of viral RNA from the original sample, it is subjected to amplification via recombinase polymerase amplification (RPA). Subsequently, in the Cas13a detection phase, the target amplicon was introduced to the Cas13a/crRNA complex. Upon binding to the SARS-CoV-2 gene, Cas13a will be activated and cleaved on surrounding non-target reporter molecules. In conclusion, the single-stranded RNA (ssRNA) reporter molecules are coupled to fluorescent molecules, allowing for straightforward observation of the cleavage event. While the Cas13a-based detection system is primarily employed for the identification of RNA viruses, it can also be utilized for the detection of DNA molecules through the addition of the T7 RNA polymerase reaction, which converts DNA into RNA molecules that are recognizable by the Cas13a element.

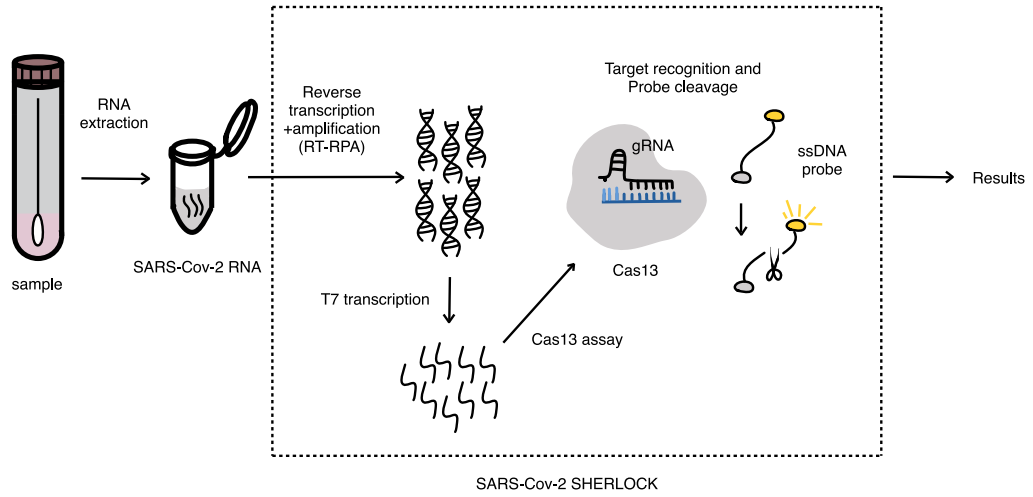

Figure 1. SHERLOCK (Specific High-sensitivity Enzymatic Reporter UnLOCKing) system

The research developed an enhanced Cas13-based genetic test for SARS-CoV-2. This system is a rapid and sensitive assay that can be performed without the need for laboratory equipment. The enhanced methodology can be employed to ascertain the presence of the SARS-CoV-2 gene in clinical specimens and can be accomplished in less than an hour with a straightforward laboratory apparatus [5].

2.2.2. DETECTR (DNA Endonuclease-Targeted CRISPR Trans Reporter) system

The DETECTR technology is a nucleic acid assay that exploits the characteristics of the Cas12a: CRISPR-Cas 12-related technology. This technology can be employed as a tool for the real-time

DOI: 10.54254/2753-8818/60/20241417

detection of SARS-CoV-2. The complete process of the assay utilising the Cas12 detection platform is shown in Figure 2. The principal processes encompass reverse transcription (RT) amplification, Cas12 reaction, and signal readout. Following the isolation of viral RNA from the sample, the target Reverse transcription and amplification are used to RNA. Subsequently, the amplified DNA is introduced to the Cas12a/guide RNA complex during the Cas12a reaction phase. Upon binding to the target DNA, Cas12a incidentally cleaves nearby non-target reporter genes. Fluorophore quencher (FQ) or fluorophore biotin (FB)-labelled ssDNA is typically employed as the reporter. Once the non-target ssDNA reporter has been cleaved, the signal can be read using either fluorescence or colorimetric reactions. Furthermore, the DETECTR assay can combine thermostable amplification with CRISPR/Cas detection. The assay time can be output by qualitative and visual lateral flow analysis, enabling rapid and visual SARS-CoV-2 nucleic acid detection.

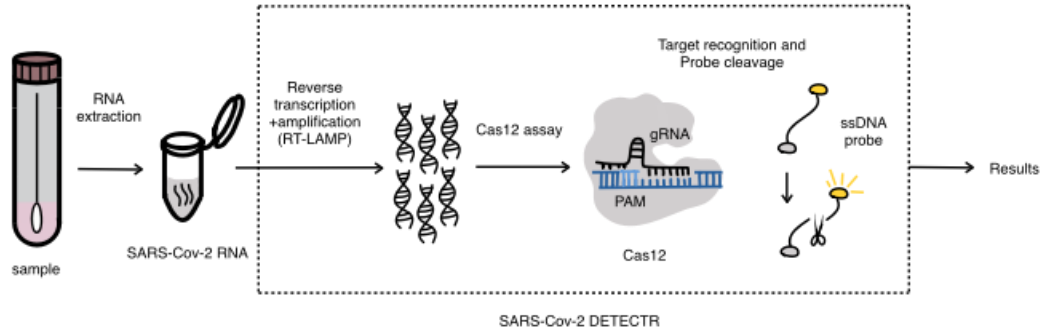

Figure 2. DETECTR (DNA Endonuclease-Targeted CRISPR Trans Reporter) system

In less than 40 minutes, the researchers were able to identify SARS-CoV-2 RNA from patient respiratory swab samples using the DETECTR approach. This method was created by integrating the CRISPR/Cas12aDETECTR system with isothermal amplification technology.

Similarly, effective CRISPR diagnostics indicate that the DETECTR diagnostic system may be a rapid, low-cost, and simple alternative to PCR. A comparison was conducted between the DETECTR method and real time fluorescence quantita- tive PCR(qRT-PCR) in the diagnosis of SARS-CoV-2 in 378 individuals, resulting in a 95% relationship. Furthermore, the DETECTR system for diagnostics can be employed to diagnose SARS-CoV-2 without the necessity for specialized equipment. Additionally, the technique demonstrated 100% specificity in identifying SARS-CoV-2. This CRISPR diagnostic system demonstrates how the DETECTR diagnostic system can be used as a simple, affordable, and quick substitute for qRT-PCR without sacrificing the molecular assay's sensitivity or specificity.

2.2.3. Other emerging CRISPR-based diagnostic tools

Although many studies reported the use of Cas12a, an alternative assay employs Cas12b. In the Cas12b-based system, RNA is initially extracted from the sample and amplified via recombinase-assisted amplification following reverse transcription to produce cDNA. The amplification product is then combined with the Cas12b complex in a test tube, allowing the modified crRNA to bind to the SARS-CoV-2 gene and cleave the reporter molecule with fluorophore, which is then visualized under blue LED light or UV light. The reaction is completed within 10-30 minutes at a temperature of 37-42°C, and is highly sensitive with a detection limit of approximately 10 copies/uL [6].

3. Machine Learning Approaches in COVID-19 Diagnostics

A significant area of research is the utilization of machine learning (ML) to comprehend and combat the COVID-19. This is currently a highly active area of research. The potential of ML in diagnostics can be categorized into two distinct areas: the utilization of readily available clinical and laboratory data to diagnose COVID-19 and to predict the risk of death and severity; and the utilization of machine-learning algorithms to design primers to predict new coronaviruses with their possible new

DOI: 10.54254/2753-8818/60/20241417

mutations. Discussions pertaining to the utilization of machine learning have focused on the characteristics of the algorithms employed, the composition of the training datasets, and the selection of features.

3.1. Introduction to Machine Learning

3.1.1. Basic principles of machine learning

Machine learning is a technique for enabling computers to learn from given data. When one wishes to analyse datasets that are too large and too extensive, this can affect the validity of the study. In these situations, the machine must automate the learning approaches for data analysis in order to create a reliable and effective channel for the derivation of the results. When the procedure is appropriately applied to biological experiment data, the outcome is frequently more desirable.

3.1.2. Types of machine learning: supervised, unsupervised, reinforcement learning

Supervised learning is a process whereby a machine learns to generate inputs and outputs based on a set of trained input-output samples. Supervised learning is the process of fitting a model to labelled data (or a subset of data), where there are some fundamental true properties, which are often measured or assigned through research conducted by humans.

Conversely, unsupervised learning methods can pick up knowledge by recognizing characteristics in unlabeled data.

Semi-supervised learning is a combination of supervised and unsupervised learning processes that contain a small amount of labelled data with An enormous quantity of unlabeled data. This can improve performance in situations where access to labelled data is costly.

3.2. Machine Learning Algorithms for COVID-19 Diagnostics

The early detection of any disease, whether communicable or non-communicable, represents a crucial step in the process of saving lives through the provision of early treatment. The implementation of rapid diagnostic and screening processes can assist in the prevention of the spread of pandemics such as SARS-CoV-2, while also being cost-effective and expediting related diagnoses. The advent of healthcare expert systems has facilitated a paradigm shift in the identification, screening, and management of SARS-CoV-2 carriers, surpassing the limitations of conventional methods.

Machine learning (ML) and artificial intelligence (AI) have been instrumental in advancing disease diagnostics and screening processes. Radio-imaging techniques, including computed tomography (CT), x-rays, and information from clinical blood samples, are employed to identify patients. Radiological images can be utilized by healthcare professionals as a routine tool to enhance traditional diagnosis and screening.

3.3. Applications of Machine Learning in Diagnostics

In this regard, a study has been conducted utilizing deep convolutional networks to propose a new model to diagnose SARS-CoV-2 in a fast and efficient way, showing the potential of AI and ML tools. To improve the accuracy of Covid-19 diagnostics, a novel model for automated Covid-19 identification founded on deep learning algorithms was created recently. The developed model used original chest radiograph images of 127 infected patients, of which 500 undetected and 500 pneumonia cases were recorded. The performance accuracy was significant, 98.08% for binary class and 87.02% for multiclass. Demonstrates the important role of expert systems in assisting in a rapid and accurate screening process [7].

4. Integration of CRISPR/Cas and Machine Learning

The CRISPR/Cas system is a potent gene editing tool. CRISPR/Cas system developers and users have benefited from current advancements in the area of gene editing involving the adaption and advancement of machine learning techniques. The experimental load of optimizing the design of

DOI: 10.54254/2753-8818/60/20241417

sgRNAs for a particular gene editing task is lessened when machine learning (ML) models are used to predict the editing outcome of a given sgRNA. Better gene editing tools can be created more easily since ML models can also predict the structure of proteins and offer a framework for directed evolution.

4.1. Conceptual Framework for Integration: the combination of CRISPR/Cas and ML

CRISPR Cas is a really unique technique that has the potential to change many industries, like agriculture, medicine, and biotechnology. This skill does not, however, come without limitations. One such disadvantage is the possibility of unforeseen changes (off-target), which emphasizes the need for precise forecasting and mitigating techniques. The utilisation of machine learning in applications has enhanced the capacity for off-target prediction. However, these studies have frequently encountered difficulties in achieving the desired precision-recall trade-off, thereby limiting their efficacy and failing to provide satisfactory explanations for the intricate decision-making processes of their models. However, current ML models related to gene editing are sometimes susceptible to confirmation bias due to the composition of training datasets, which restricts their utility. It is imperative to develop more robust models and extend the application of ML to other facets of CRISPR/Cas gene editing [8].

4.2. Benefits of integration: Speed, accuracy, and scalability

Beyond what typical linear models can do, deep learning in machine learning can be used to identify complicated non-linear relationships in data. Its network is made up of several interconnected layers that perform a number of nonlinear changes on the input data before producing predictions as an output. Given the complexity of the sequence-activity link, this technique makes complex patterns in the input data visible, a fact that is especially important for gRNA activity prediction.

Several studies have been conducted with the objective of predicting the outcome of CRISPR-Cas9 edits. Each study has employed a distinct approach and has yielded insights that are unique to that study. For example, the Apindel model [9], a framework combining attentional and bi-directional long short-term memory (BiLSTM) mechanisms, was first given by Liu. The model demonstrated superior performance in terms of prediction details and accuracy compared to previous models. Similarly, Li W. and others proposed Sequence Generative Adversarial Nets(SeqGAN), a novel model that integrates convolutional neural networks (CNN) and sequence generation adversarial networks to enhance the prediction of CRISPR off-target cutting sites[10].

4.3. Case Studies of Integrated Systems in COVID-19 diagnostics

Ameen developed a deep learning tool based on the combination of Support Vector Regression (SVR) and CNN, which can assist in the selection of appropriate gRNAs to successfully target specific regions on the SARS-CoV-2 gene using the Cas12a system. Another deep learning model employs CNN to predict active and inactive gRNAs[11]. In a separate study, Metsky and Freije devised broad array of experimental methods and test designs for a CRISPR-based diagnostic system that may prove advantageous for long-term monitoring. The designs described were employed to detect 67 distinct virus species and subspecies, including SARS-CoV-2. Metsky's approach employs machine learning algorithms to develop molecular assays, with the objective of enhancing the accuracy of virus species identification through diagnostics. The approach offers advances in three areas: (1) prediction of enzyme activity for diagnostics; (2) optimal integration of viral variation into the design of diagnostics; and (3) rapid design of diagnostics at large scale [12].

Wei developed the Cas13 platform for high-throughput phenotypic screening and elucidated the design principles that support its RNA targeting effectiveness. The researchers compared various machine learning algorithms and linear regression in order to train a CNN that could predict gRNA activity based on gRNA sequences. The final CNN model demonstrated an accuracy of over 90% in predicting active gRNAs on a test dataset. Additionally, the study developed a design tool that could be used to create an activity guide for detecting RNAs using the Cas13 system[13].This model, which incorporates machine learning for guideRNA prediction, postulates that regular runs can be adapted to

DOI: 10.54254/2753-8818/60/20241417

thousands of viruses to maintain the design's reflection of evolutionary processes. Adapt's design demonstrates efficacy for known viruses and is even useful for new viruses not yet known during the design process. Nevertheless, the performance of this approach may be limited when applied to novel viruses due to potential sampling bias in the genome data.

it is essential that the body of knowledge in this area be updated often. This fosters a collaborative atmosphere that encourages innovation in addition to making it easier for scientists to share findings with one another. By doing thus, it has contributed significantly to the development of a scientific ecosystem that can successfully address new problems and seize possibilities.

4.4. Challenges and Limitations

Data quality and availability issues

A common characteristic of CRISPR-Cas investigations was the virtually universal use of distinct datasets, with each study typically utilizing an author-specific dataset. As a result, the measures used differ greatly, spanning from cosine similarity to accuracy, AUC, and log-likelihood values. Using public datasets for numerical comparisons is challenging due to the diversity of these indicators. Numerical performance comparisons of current approaches face a substantial hurdle due to tLack of harmonised reference data sets and comparable metrics in the field. This review takes into account this intrinsic restriction of existing CRISPR-Cas research and suggests that future studies aiming at advancing the field pay heed to it.

5. Conclusion

By altering the crRNA sequence, it is feasible to target other gene sequences due to the considerable programmability and flexibility of CRISPR. Because of this, the CRISPR system is able to rapidly adjust in response to the majority of novel variants resulting from recently found mutations. Machine learning, on the other hand, is a relatively new technology that has made significant innovations in various fields. Among these, attention mechanisms represent a recent innovation. These techniques are derived from natural language processing, where they enhance the interpretability of deep learning models by concentrating on important segments of the input sequence. When combined, these patterns show how quickly deep learning techniques have advanced in the field of gRNA activity prediction. These developments illustrate the growing maturity of the field, while simultaneously identifying a multitude of untapped avenues for further exploration. In conjunction with the continuous development of deep learning architectures and techniques, these developments present numerous potential paths toward improving the accuracy and effectiveness of gRNA activity prediction in the future. The CRISPR-Cas system has great potential in the future. In the fight against SARS-CoV-2, CRISPR-Cas technology has demonstrated its powerful diagnostic capabilities. It is expected that assays based on machine learning and CRISPR will be applied more broadly to aid in the identification of further newly discovered viruses.

References

[1] Li, Q., Guan, X., Wu, P., et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *%J New England journal of medicine, 382*(13), 1199-1207. https://doi.org/10.1056/NEJMoa2001316

[2] Curti, L. A., Primost, I., Valla, S., et al. (2021). Evaluation of a lyophilized CRISPR-Cas12 assay for a sensitive, specific, and rapid detection of SARS-CoV-2. *Llases, María Eugenia %J Viruses, 13*(3), 420. https://doi.org/10.3390/v13030420

[3] Greener, J. G., Kandathil, S. M., Moffat, L., et al. (2022). A guide to machine learning for biologists. *Nat Rev Mol Cell Biol, 23*(1), 40-55. https://doi.org/10.1038/s41580-021-00407-0

[4] Jinek, M., Chylinski, K., Fonfara, I., et al. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *%J science, 337*(6096), 816-821. https://doi.org/10.1126/science.1225829

[5] Ooi, K. H., Tay, J. W. D., Teo, S. Y., et al. (2020). A CRISPR-based SARS-CoV-2 diagnostic assay that is robust against viral evolution and RNA editing. *%J BioRxiv*. https://doi.org/10.1101/2020.07.03.185850

[6] Guo, L., Sun, X., Wang, X., et al. (2020). SARS-CoV-2 detection with CRISPR diagnostics. *%J Cell discovery, 6*(1), 34. https://doi.org/10.1038/s41421-020-0174-y

[7] Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals, 139*, 110059. https://doi.org/10.1016/j.chaos.2020.110059

[8] Lee, M. (2023). Deep learning in CRISPR-Cas systems: a review of recent studies. *Front Bioeng Biotechnol, 11*, 1226182. https://doi.org/10.3389/fbioe.2023.1226182

[9] Liu, X., Wang, S., & Ai, D. (2022). Predicting CRISPR/Cas9 repair outcomes by attention-based deep learning framework. *%J Cells, 11*(11), 1847. https://doi.org/10.3390/cells11111847

[10] Li, W., Wang, X.-B., & Xu, Y. (2022). Recognition of CRISPR off-target cleavage sites with SeqGAN. *%J Current Bioinformatics, 17*(1), 101-107. https://doi.org/10.2174/1574893616666210727162650

[11] Ameen, Z. S. i., Ozsoz, M., Mubarak, A. S., et al. (2021). C-SVR Crispr: Prediction of CRISPR/Cas12 guideRNA activity using deep learning models. *%J Alexandria Engineering Journal, 60*(4), 3501-3508. https://doi.org/10.1016/j.aej.2021.02.007

[12] Metsky, H. C., Welch, N. L., Pillai, P. P., et al. (2022). Designing sensitive viral diagnostics with machine learning. *Nat Biotechnol, 40*(7), 1123-1131. https://doi.org/10.1038/s41587-022-01213-5

[13] Wei, J., Lotfy, P., Faizi, K., et al. (2021). Deep learning of Cas13 guide activity from high-throughput gene essentiality screening. *%J BioRxiv*. https://doi.org/10.1101/2021.09.14.460134