CRISPR-Cas9 Guide RNA Designer using Python

Arya Hariharan   
*Department of Computer Science and Engineering*   
*RV College of Engineering*Bengaluru, India  
aryahariharan.cs22@rvce.edu.in

B H Abhisha  
*Department of Computer Science and Engineering*   
*RV College of Engineering*Bengaluru, India  
bhabhisha.cs22@rvce.edu.in

Narendra Kumar S  
*Department of Biotechnology*   
*RV College of Engineering*Bengaluru, India  
narendraks@rvce.edu.in  
Bhumika K  
*Department of Computer Science and Engineering*   
*RV College of Engineering*Bengaluru, India  
bhumikak.cs22@rvce.edu.in

*Abstract*—The CRISPR-Cas9 gene editing, based on a bacterial immune mechanism, offers precise DNA editing techniques for gene therapy, disease treatment, and improved agricultural crops. This paper reviews the various aspects of the technology and outlines how a bioinformatics tool that combines sequence retrieval, characteristics analysis, and user interaction can be used create a systematic framework for building guide RNAs for CRISPR-Cas9 gene editing. The tool, built using Python and Streamlit, allows users to enter in the DNA sequence of choice, and automatically extract guide RNAs that can potentially be used for gene editing. Then, based on user inputs, it then ranks the extracted sequences from least compatible to most compatible for the user's purpose.

Keywords—CRISPR-Cas9, gene editing, guide RNA, Python

# Introduction

The CRISPR-Cas9 mechanism, derived from bacteria's adaptive immune defenses, is a significant advancement in genetic engineering. This genome-editing technique allows for precise and targeted modifications to DNA sequences, enabling researchers to edit genes with control and proficiency. CRISPR-Cas9 functions through the use of a guide RNA to convey the Cas9 protein to a precise site in the genome, initiating a double-strand break at that specific location. The cell's intrinsic repair frameworks attempt to mend the break, potentially leading to the introduction, evacuation, or modification of hereditary material.

Since its discovery nearly a decade ago in Californian archives, the bacterial CRISPR-Cas9 system has revolutionized the fields of medicine, agriculture, and basic science due to its unprecedented precision in editing genomes. In the medical domain, it has the potential to correct mutations responsible for cystic fibrosis, sickle cell disease, and Duchenne muscular dystrophy at their genetic roots. In agriculture, CRISPR technology has enabled scientists to genetically enhance crops in desirable traits such as heightened resistance to pests and pathogens, augmented yields, and more nutritious nutritional profiles.

However, CRISPR-Cas9 practical use requires advanced methods for feature evaluation, guide RNA (gRNA) design, and sequence analysis. The effectiveness of CRISPR-Cas9 research depends on the design of gRNAs that maximize on-target efficiency while minimizing off-target consequences. Additionally, sequence properties including GC content, secondary structures, and possible off-target locations must be analyzed. To address these demands, this paper proposes the development of CRISPR-Cas9 guide RNA designer tool. This tool facilitates the CRISPR-Cas9-based genome editing workflow, from gRNA identification to gRNA feature analysis. By utilizing Streamlit's interactive web interface and Python's computational analysis capabilities, the application opens up advanced bioinformatics to a wider audience.

# Literature Survey

Gene editing is the alteration of genetic material in organisms to induce certain specific traits or characteristics in an organism. Gene editing technologies utilise the concept of targeted nucleases to manipulate genomic sequences in a site-specific manner. It allows to rapidly introduce modifications in genomic sequences for a broad range of cells and organisms.

The clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR associated protein 9 (Cas9) based system is one such gene editing technology. The use of CRISPR-Cas9 in gene editing was understood based on the role it played in the adaptive immunity of bacterial cells [1]. When a bacterium encounters an invasive virus or plasmid, it captures short sequences of its DNA. It then uses RNA-guided Cas9 proteins to cleave its own DNA and insert these sequences of foreign DNA into the genome at specific locations called the CRISPR locus. These sequences are called spacers and are interspersed with short repeated sequences called repeats. The CRISPR locus is transcribed into a long precursor RNA molecule which is then processed into shorter CRISPR RNAs (cRNAs), each containing a sequence complementary to a segment of the foreign DNA. The crRNA forms a complex with another RNA molecule called the trans-activating crRNA (tracrRNA). The tracrRNA is partially complementary to the repeat regions in the crRNA.The crRNA-tracrRNA complex then binds to the Cas9 protein, forming an active CRISPR-Cas9 complex, which is an endonuclease and thus, can cut DNA. The crRNA guides the complex to a specific sequence in the foreign DNA that matches the spacer sequence (now part of the crRNA). This sequence in the foreign DNA is called the protospacer. Once the CRISPR-Cas9 complex binds to the matching sequence in the foreign DNA, the Cas9 protein makes a double-stranded cut in the DNA at that location, thereby disabling the foreign DNA [2-4].

This concept has since been replicated and simplified to be applied in genome engineering, leading to the invention of the CRISPR-Cas9 genome editing technology. It involves two essential components: a guide RNA sequence that acts as a template for identifying the location in the genome sequences where the editing has to be done and the Cas9 protein which, being an endonuclease, can cause a double stranded DNA break for gene modification. The identification of the target site is done entirely by the guide RNA, removing the need for engineering new proteins every time a new target site is identified. Thus, it is one of the most flexible and user-friendly methods for genome editing [5].

CRISPR-Cas9 based gene editing techniques can be potentially employed in the correction of genetic disorders by identifying and modifying the disease causing gene, treatment of infectious diseases like HIV and modification of patient T-cells to cure diseases [6]. Reference [7] describes the first demonstration of using CRISPR-Cas9 for targeted gene editing in eukaryotic cells. Expanding on this, Reference [8] focuses on Cpf1, an alternative Cas9 enzyme with different cutting properties. Cpf1's unique features, such as producing staggered cuts in DNA rather than the blunt cuts made by Cas9.

The CRISPR/Cas system has provided versatile tools for manipulating, detecting, imaging, and annotating specific DNA or RNA sequences, with Cas proteins like Cas9, Cas12, Cas13, and Cas14 facilitating genome engineering and advancing genetic research across diverse organisms. CRISPR/Cas systems can also be used in the detection of viruses such as SARS-CoV-2 [9]. CRISPR/Cas13a has been employed to engineer interference with the Turnip Mosaic Virus (TuMV) in plants, demonstrating effective targeting of RNA virus sequences, particularly the HC-Pro and GFP regions, and offering a novel RNA-guided mechanism for virus interference and RNA manipulation in plant systems [10]. CRISPR/Cas genome editing technology has also been utilized for targeted mutagenesis in rice, with methodologies encompassing guide RNA design, vector construction, transformation protocols, and molecular analysis for precise genetic alterations aimed at enhancing desired traits [11]. CRISPR/Cas9 technology has recently been applied to Medicago sativa (alfalfa) to improve pathogen resistance, involving guide RNA design, binary CRISPR vector construction, Agrobacterium-mediated transformation, and molecular assessments for transgene and edit identification, facilitating targeted genetic alterations for enhanced agronomic traits [12].

Reference [13] investigates methods for optimizing CRISPR-Cas9 for gene editing in mammalian cells, discussing strategies to improve the system's efficiency and delivery. This includes enhancing the targeting accuracy of the Cas9 enzyme and increasing the efficiency of CRISPR delivery mechanisms into cells. In Reference [14], the use of CRISPR-Cas9 for gene editing in fruit flies (Drosophila) is demonstrated, showcasing the system's effectiveness in studying the function of specific DNA elements. This research highlights the utility of CRISPR-Cas9 in functional genomics and its potential for uncovering gene functions in model organisms.

The efficacy of the technology in the applications described above is largely dependent on the quality of the guide RNAs used to identify cleavage points within the genome sequences. Poorly designed guide RNA can lead to improper binding, low cleavage efficiency and off-target binding. Thus, guide RNA sequences have to be engineered carefully, keeping several factors in mind. The efficacy is further determined by what is trying to be achieved. Guide RNA designs vary for gene knockout, base editing and modulation of the gene expression.

Target recognition by guide RNA relies on the presence of a protospacer adjacent motif (PAM) sequence located 2 to 5 base pairs away from the target DNA. The Cas proteins bind to the PAM sequences in gene sequence, and then proceeds with the cleavage.  There are different methods to identify the PAM sequences, based on either in-silico, in-vivo or in-vitro techniques. These approaches include, but are not limited to, alignment of protospacers, plasmid depletion assays and PAM screening achieved by NOT-gate repression [15]. The PAM sequence is required to be at 3’ end of the target sequence. This sequence also varies across different bacterial species. The PAM for the commonly utilized Cas9 from Streptococcus pyogenes is usually NGG, with no nucleotide bias seen in the first position [16].

A guide sequence of 20 nucleotides is typically used in CRISPR-Cas9 based systems. However, guide RNA sequences of varying length do have an impact on the efficiency of the system. Studies have found that 17 or above base pair sequence lengths generally work well in directing the Cas9 protein to the correct location, where 15 or lesser base pair sequence lengths resulted in poorer performance [17]. Position-specific nucleotide composition also affects the quality of the guide RNA. Several positions have preferable nucleotides which can lead to greater stability and on-target activity, with nucleotide characteristics at positions 16-20, called the seed region, showing the greatest impact on efficiency [18]. Guanine or adenine is highly preferable at position 19 and 20,  and cytosine is highly preferred at position 16 and 18. Position 17 does not show any clear preference for a specific nucleotide [19-23].

GC content refers to the percentage of a gene, gene region, chromosome, or genome that is guanine (G) or cytosine (C) as opposed to adenine (A) or thymine (T). GC content is also seen to have an impact on the thermodynamic stability of gRNA contributed by greater extents of hydrogen bonding in the GC pairs. GC content must typically be in the range of 40- 60%, and gRNAs with GC content out of this range were inefficient [24].

Off-target effects, or the introduction of unintended, undesirable, or even harmful changes to the genome, are also a significant factor and cause of worry in the uses of the CRISPR/Cas9 system. Different methods exist for detecting the number of potential off-targets of a given gRNA sequence. In silico methods are mainly based on alignment based or scoring based models. Experimental detection methods can also be used based on cell-free detection, culture-free detection and in vivo detection. Experimental methods tend to be more expensive compared to in silico methods but can be more accurate due to experimental validation [25].

CRISPR-Cas9 gene editing technology is a highly promising technology in the field of gene engineering. Identifying guide RNA sequences is a key and primary requirement for the process and there are several factors that influence the effectiveness of the sequence. Manually picking out possible sequences and computing the effectiveness is a tedious task, especially as gene sequences become longer in length. Automation of this process can aid greatly in helping researchers detect all possible gRNA sequences and understand the relative effectiveness of these sequences on a set of fixed parameters. In the coming sections, the design and implementation of such an automation tool is discussed.

# Methodology

This section details the methodology followed in building the tool for detecting possible guide RNA sequences in any given gene sequences. The goal was to design an easy-to-use tool that allows users to enter in a gene sequence of interest and find possible gRNA sequences that they can use for CRISPR-Cas9 gene editing. In order to do so, the Python programming language was used for parsing input sequences to search for possible gRNA sequences. The logic of evaluating the detected sequences was also implemented in Python. In order to create a simple yet effective user interface, the Streamlit library of Python was used. The methodology is represented in Figure 1.

In order to allow greater flexibility, the tool was designed to accept two types of input. The user can enter the genome sequence for which the guide RNAs have to be found either by pasting the sequences inside a text field or by uploading a FASTA file. For easy parsing, the enter FASTA file’s content is slightly modified by removing newlines and creating a single sentence sequence.

Two modules have been written for the two primary functions of the tool. The first module is responsible for detecting all possible gRNA sequences from the given input. The tool is

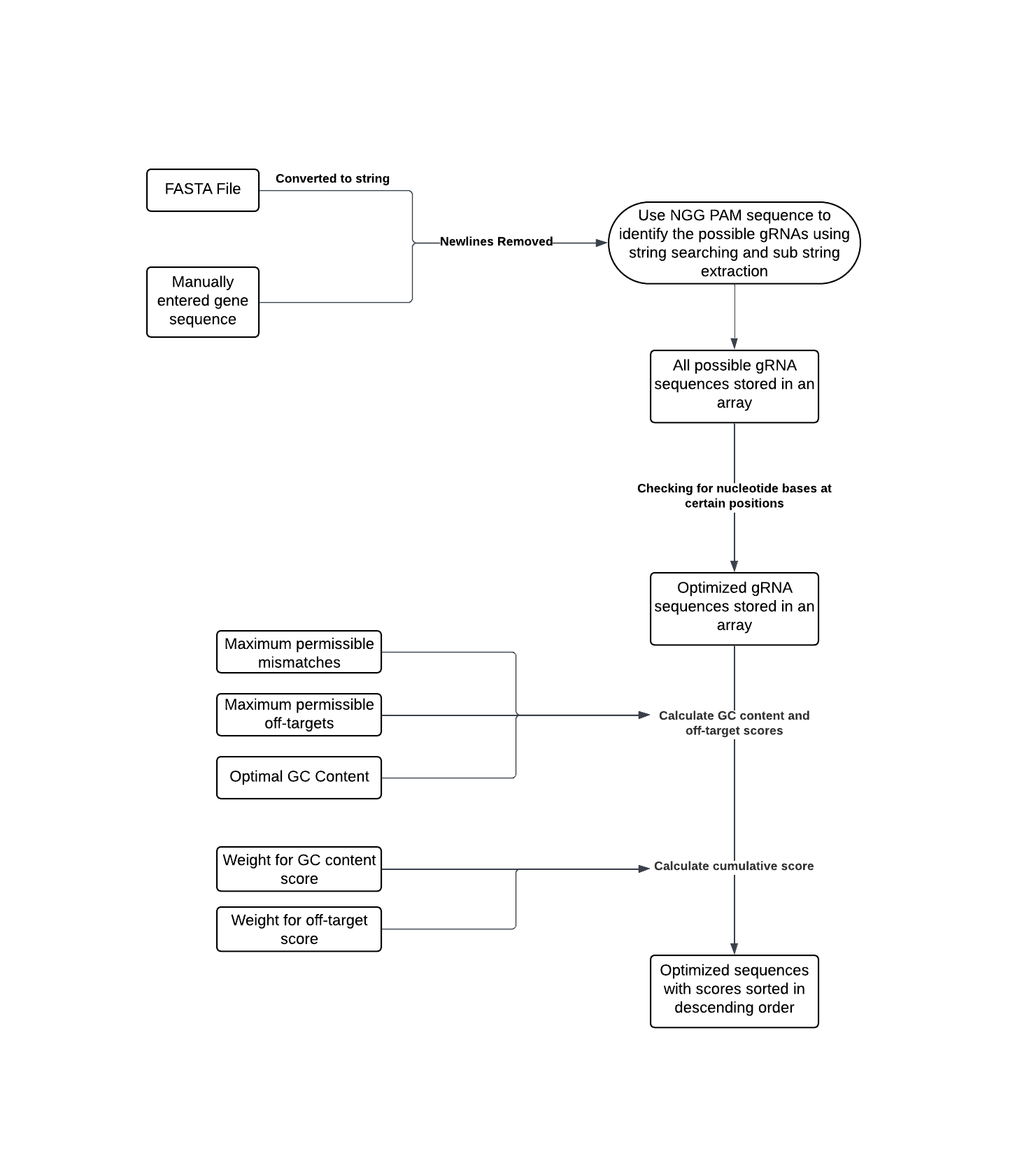


Figure 1

designed specifically for Cas9 proteins derived from Streptococcus pyogenes and thus, considers the PAM sequence to be NGG, with no bias for the first base of the PAM sequence. The input is considered to be of string type. In case of FASTA file input, the content of the file is converted into string type by using the StringIO function of the ‘io’ module in Python. The module functions are then invoked to identify all possible sequences of length 20 that are suffixed by an NGG protospacer, which is extracted by using the sub-string extraction function offered implicitly by Python. These sequences are identified sequentially as the entire sequence is parsed through, and the identified gRNA sequences are stored in a list. A condition is placed such that only uniquely encountered sequences are placed in the list. Repeated sequences, if identified, are not stored. Here, optimised sequences are also checked for, by considering if certain bases are present at preferred locations within the sequence, as discussed in the previous section [18-23]. It is checked whether cytosine is present at position 16 and 18 and whether adenine or guanine is present at positions 19 and 20. Sequences satisfying these conditions are considered as optimized sequences and are also placed within another list along with other optimized sequences.

Following the identification of sequences, the second module aims to understand the features of the optimized sequences identified. The features include number of possible off-targets for a given number of mismatches and GC content. The user is allowed to choose the maximum possible number of mismatches in off-targets, the maximum permissible off-targets and optimal GC content for their use case. Based on the maximum mismatches allowed, the number possible off-target sequences are identified through string matching. Off-targets with mismatches equal to or lesser in number than the value entered by the user are counted. A score is assigned based on this count. If the off-target count is greater than the maximum number of off-targets entered by the user, then the off-target score of the sequence is set as 0. Else, a score less than or equal to 10 is allocated, following Equation 1.

The GC content of the sequences is also calculated with a simple formula given by Equation 2.

A score is also allocated to the calculated GC content based on Equation 3.

The user can then apply weights to each of these scores. The weights are applied so as to calculate a cumulative score for each sequence, that considers both GC content as well number of off-targets. The weights are used to assign a priority to the different features, with higher weights implying higher priority to a feature. The weights are multiplied with the scores calculated previously and the resultants are added to get the cumulative score for a sequence, given by Equation 4. The same is done for all possible sequences and finally, the optimized sequences are ranked in a descending order based on their cumulative scores and are displayed to the user. Higher cumulative scores imply better optimization and affinity to the features described by the user.

# Results and Discussions

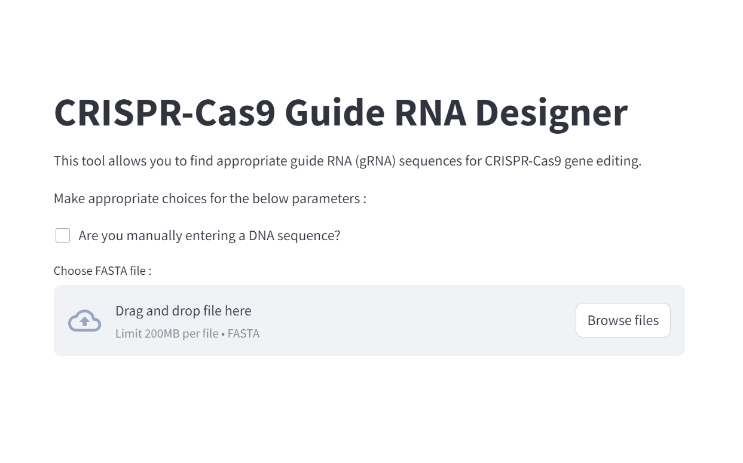
 This section expands on the results achieved by implementing the method described in the above section. Using Streamlit, the user interface as shown in Figure 1 is created.

Figure 2

The default option is for the user to choose a FASTA file from their local device. On checking the checkbox, the user is allowed to manually enter the DNA sequence using a text field as seen in Figure 3.

After entering or uploading the sequence, the number of possible gRNA sequences along with the number of optimized gRNA sequences that have been detected is displayed to the user. The user is then allowed to set the values for the maximum permissible mismatches, off-targets and ideal GC content percentage, as shown in Figure 4.

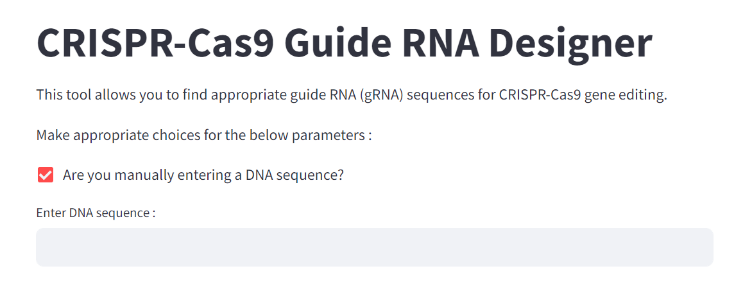


Figure 3

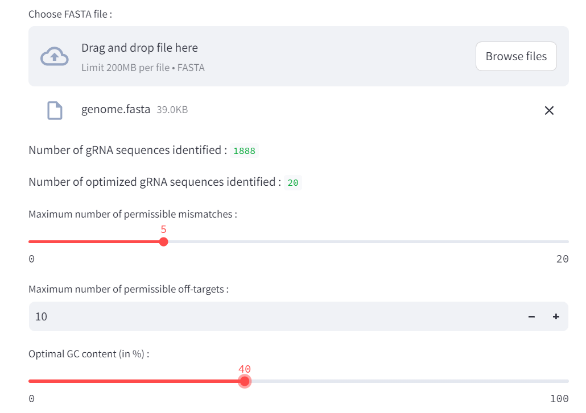


Figure 4

Setting a non-zero value for GC content is a necessary requirement. Else, the score calculation will not be done. After this, the users are prompted to add in the weights. The weights must be an integral value that add up to 10, as shown in Figure 5. Internally, the value out of 10 is converted to a value out of 1 by dividing it by 10. If the weights do not add up, a suitable error message is displayed, as in Figure 6.

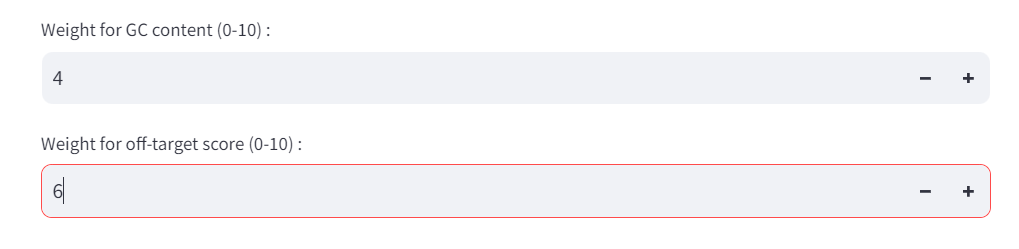


Figure 5

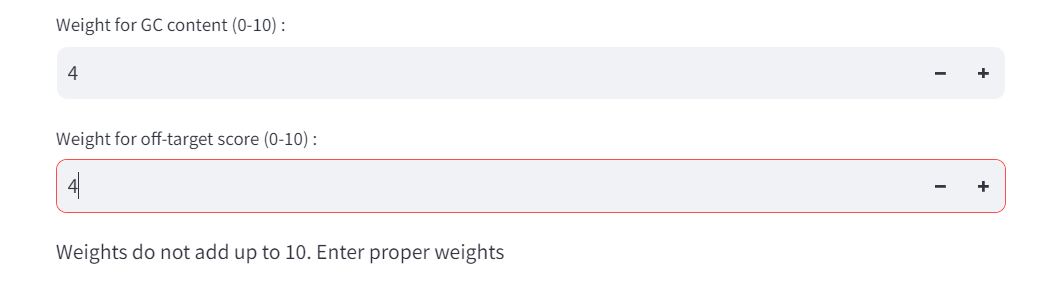


Figure 6

Once entered correctly, the calculation for cumulative score is done and the optimized gRNA scores sorted as per their scores are displayed in an expandable format to the user. Expanding displays the sequences and their scores as in Figure 7.

Thus, the tool accurately detects the required sequences and displays them to the user. It can be utilized by researchers to identify potential gRNA sequences for CRISPR-Cas9 gene editing.

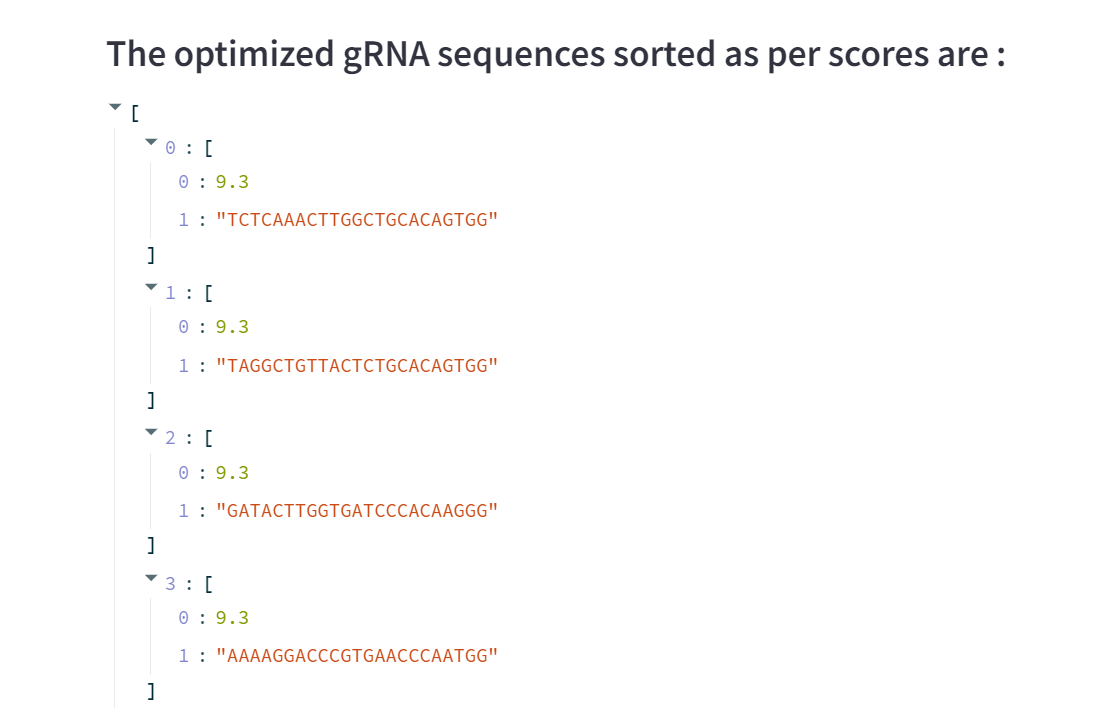


Figure 7

The scope of the tool can be greatly expanded. The tool can be integrated with existing databases to allow users to directly choose the gene sequence of interest from the database, by entering abbreviated names of the gene sequence. It can also be modified to include greater variety in terms of PAM sequences. Parallel processing can be used to speed up the time of detecting and evaluating the gRNA sequences. Deep learning models trained on gRNA data could also be employed as an alternative method for understanding the efficiency of the gRNA sequence.

##### References

1. Marraffini, L., Sontheimer, E. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11, 181–190 (2010). <https://doi.org/10.1038/nrg2749>
2. Wiedenheft, B., Sternberg, S. & Doudna, J. RNA-guided genetic silencing systems in bacteria and archaea. Nature 482, 331–338 (2012). <https://doi.org/10.1038/nature10886>
3. Sorek R, Lawrence CM, Wiedenheft B. CRISPR-mediated adaptive immune systems in bacteria and archaea. Annual review of biochemistry. 2013 Jun 2;82(1):237-66.
4. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. science. 2012 Aug 17;337(6096):816-21.
5. Gaj, Thomas et al. “Genome-Editing Technologies: Principles and Applications.” Cold Spring Harbor perspectives in biology vol. 8,12 a023754. 1 Dec. 2016, doi:10.1101/cshperspect.a023754
6. Redman, Melody et al. “What is CRISPR/Cas9?.” Archives of disease in childhood. Education and practice edition vol. 101,4 (2016): 213-5. doi:10.1136/archdischild-2016-310459
7. Anders, C., Niewoehner, O., Duerst, A. et al. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513, 569–573 (2014). <https://doi.org/10.1038/nature13579>
8. Arribere JA, Bell RT, Fu BX, Artiles KL, Hartman PS, Fire AZ. Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in Caenorhabditis elegans. Genetics. 2014 Nov;198(3):837-46. doi: 10.1534/genetics.114.169730. Epub 2014 Aug 26. PMID: 25161212; PMCID: PMC4224173.
9. Hillary VE, Ceasar SA. A Review on the Mechanism and Applications of CRISPR/Cas9/Cas12/Cas13/Cas14 Proteins Utilized for Genome Engineering. Mol Biotechnol. 2023 Mar;65(3):311-325. doi: 10.1007/s12033-022-00567-0. Epub 2022 Sep 27. PMID: 36163606; PMCID: PMC9512960.
10. Aman R, Ali Z, Butt H, Mahas A, Aljedaani F, Khan MZ, Ding S, Mahfouz M. RNA virus interference via CRISPR/Cas13a system in plants. Genome Biol. 2018 Jan 4;19(1):1. doi: 10.1186/s13059-017-1381-1. PMID: 29301551; PMCID: PMC5755456.
11. Xu R, Wei P, Yang J. Use of CRISPR/Cas Genome Editing Technology for Targeted Mutagenesis in Rice. Methods Mol Biol. 2017;1498:33-40. doi: 10.1007/978-1-4939-6472-7\_3. PMID: 27709567.
12. Meng X, Hu X, Liu Q, Song X, Gao C, Li J, Wang K. Robust genome editing of CRISPR-Cas9 at NAG PAMs in rice. Sci China Life Sci. 2018 Jan;61(1):122-125. doi: 10.1007/s11427-017-9247-9. Epub 2017 Dec 25. PMID: 29285711.
13. David Bikard, Wenyan Jiang, Poulami Samai, Ann Hochschild, Feng Zhang, Luciano A. Marraffini, Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system, Nucleic Acids Research, Volume 41, Issue 15, 1 August 2013, Pages 7429–7437, <https://doi.org/10.1093/nar/gkt520>
14. Gagnon JA, Valen E, Thyme SB, Huang P, Ahkmetova L, Pauli A, et al. (2014) Efficient Mutagenesis by Cas9 Protein-Mediated Oligonucleotide Insertion and Large-Scale Assessment of Single-Guide RNAs. PLoS ONE 9(5): e98186. <https://doi.org/10.1371/journal.pone.0098186>
15. Gleditzsch, Daniel et al. “PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures.” RNA biology vol. 16,4 (2019): 504-517. doi:10.1080/15476286.2018.1504546
16. Wu, Xuebing et al. “Target specificity of the CRISPR-Cas9 system.” Quantitative biology (Beijing, China) vol. 2,2 (2014): 59-70. doi:10.1007/s40484-014-0030-x
17. Lv, Jiawei et al. “The length of guide RNA and target DNA heteroduplex effects on CRISPR/Cas9 mediated genome editing efficiency in porcine cells.” Journal of veterinary science vol. 20,3 (2019): e23. doi:10.4142/jvs.2019.20.e23
18. Wu, X., Scott, D., Kriz, A. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat Biotechnol* 32, 670–676 (2014). <https://doi.org/10.1038/nbt.2889>
19. Doench J.G., Hartenian E., Graham D.B., Tothova Z., Hegde M., Smith I., Sullender M., Ebert B.L., Xavier R.J., Root D.E.. Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. Nat. Biotechnol. 2014; 32:1262–1267.
20. Wong N., Liu W., Wang X.. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biol. 2015; 16:218.
21. Chari R., Mali P., Moosburner M., Church G.M.. Unraveling CRISPR–Cas9 genome engineering parameters via a library-on-library approach. Nat. Methods. 2015; 12:823–826.
22. Xu H., Xiao T., Chen C.H., Li W., Meyer C.A., Wu Q., Wu D., Cong L., Zhang F., Liu J.S.et al... Sequence determinants of improved CRISPR sgRNA design. Genome Res. 2015; 25:1147–1157.
23. Konstantakos, Vasileios et al. “CRISPR-Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning.” Nucleic acids research vol. 50,7 (2022): 3616-3637. doi:10.1093/nar/gkac192
24. Konstantakos, Vasileios et al. “CRISPR-Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning.” Nucleic acids research vol. 50,7 (2022): 3616-3637. doi:10.1093/nar/gkac192
25. Guo, Congting et al. “Off-target effects in CRISPR/Cas9 gene editing.” Frontiers in bioengineering and biotechnology vol. 11 1143157. 9 Mar. 2023, doi:10.3389/fbioe.2023.1143157