ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

Cas9 protein screening of microbial data based on biological information and VAE methods

Yiting Liu

Stevenson School, Pebble Beach, Monterey, California, USA
tinayitingliu@outlook.com

Abstract: *Gene editing technology, particularly the CRISPR-Cas9 system, has revolutionized biological research, offering vast therapeutic potential. However, challenges like off-target effects, limited targetable sequences, and DNA strand fractures impede its widespread application. To tackle these issues, a project merges bioinformatics screening and deep learning models, aiming to screen novel CRISPR-Cas9 proteins. This endeavor focuses on screening for new Cas9 proteins from Streptococcus pyogenes and Archaea genomes. The process involves genome searches and predictions using a Variational Autoencoder (VAE) model. Preliminary validation with Nblast and constructing an evolutionary tree of protein ortholog distribution assess specificity. The goal is to discover Cas9 proteins that surpass current gene editing limitations, complementing the existing CRISPR/Cas system. This project advances gene editing therapies by presenting a comprehensive workflow combining bioinformatics and deep learning, serving as a valuable reference for future research.*

Keywords: *CRISPR-Cas9, Sequence homology, CD-search, pfam, VAE*

1. Introduction

CRISPR sequences, first discovered in E. coli, consist of unique 23-47 nucleotide repeats separated by spacers of foreign DNA. The CRISPR-Cas system, classified into Class I and II based on Cas proteins, functions as an immunity system in bacteria and archaea, and is widely used for gene editing due to its precision and efficiency. The rapid development of CRISPR-Cas, particularly the type II system with SpyCas9, has boosted its popularity in treating tumors and polygenetic diseases [1-2]. Successful applications include treating sickle cell disease (SCD) and transfusion-dependent β-thalassemia (TDT) [3], restoring glucose homeostasis in diabetic mice [4-5], enhancing tumor-fighting cells [6], and constructing disease models [7], as well as providing potential therapies for Alzheimer's [8], hypercholesterolemia, and Parkinson's-like symptoms. Despite its wide application, the CRISPR-Cas9 system still exhibits drawbacks, such as off-target effects due to misbinding of gRNA [9]. Meanwhile, the PAM sequence recognition specificity along with the double-strand cleavage property of Cas9 can limit applicability [10] and pose potential harm, necessitating the development of a more stable Cas9 variant.

In response to the existing problems in the current CRISPR-Cas9 gene editing system, this article aims to screening new CAS9 proteins from candidate bacterial genera and Archaea genome by identifying CRISPR palindromic repeats, thus providing an enhanced tool for gene editing and therapeutic usages. The screening process would be performed via two separate paths: search of existing microbial genomes and a Variational Autoencoder (VAE) deep learning algorithm prediction. For the genome search results, a preliminary evaluation on the newly discovered predictions by Nucleotic Blast (BLAST) would be conducted on the candidate outcomes to retrieve the similarity between the sequences we obtained and those published in known papers. A pylogenetic tree of distribution would be constructed for the screened results. The aim of our work is to retrieve the cas9 protein sequence genome on the existing genome, thereby providing a reference in the current stage of the gene editing system barriers and research gene editing therapy.

2. Method

2.1 JGI-IMG database

In this research, we used the JGI-Integrated Microbial Genomes (IMG) database, established in 2005.

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

IMG serves as an integrated microbial genome bank, containing data on bacteria, archaea, plasmids, and viruses for annotation and comparative analysis. It includes detailed information on microbial genomes, such as species classification, living conditions, genome length, coding genes, data quality, and research project details. As of today, IMG hosts over 77 billion entries across 200,000 databases [11].The genomic data for this experiment were downloaded from the IMG database.

2.1.1 Streptococcus

Streptococcus pyogenes is a Gram-positive, facultative anaerobic bacterial pathogen, primarily affecting children through infections like scarlet fever, pharyngitis, and impetigo. These bacteria spread via respiratory droplets, skin contact, and fomites. This species was chosen for the filtering process because the most commonly used Cas9 protein in research originates from this genus. The Type II CRISPR system and its Cas9 protein stem from the immune system of S. pyogenes, with SpyCas9 being the most widely utilized variant. Consequently, results from this genus are expected to reveal new Cas9 proteins, given ongoing research developments over a decade since the initial discovery of the widely used Streptococci-derived Cas9.

2.1.2 Neisseria

Neisseria are spehrical gram-negative cocci that appears mostly in pairs. Major pathogens from this genome could cause purulent cerebrospinal meningitis through respiratory infections, and Gonorrhea through sexual contact. Similar to Streptococci, the cas9 protein discovered in Neisseria has also been used as a mainstream gene editing tool in clinical trials. Therefore, the Neisseria genus is also a key subject of the screening process for Cas9 proteins.

2.1.3 Archaea

Archaea are widely distributed single celled prokaryotic organisms that are mostly found in extreme conditions. Archaea are found to be utilizing one specific set of CRISPR-Cas system derived from all the distinct CRISPR-Cas loci in its genome to defend themselves from archaeal viruses(e.g Metallosphaera sedula and its 30 CRISPR spacers used to defend itself from Acidianus two-tailed virus) or foreign DNA, thus it is used for this experiment as a potential Cas9 protein source.

2.1.4 Other bacterias

During the screening process of other bacteria species, it is found to be expect that the number of CAS proteins obtained might be much lesser than the two genera listed above that are confirmed to contain CAS proteins. Thus to ensure enough significant results could be obtained, the search scope for some of these bacterial species were expanded to genus or phylum level. An additional 7 categories of bacteria were selected to ensure a sufficient number of screened outcomes and the exclusivity of each Cas9 protein, including: Acidobacteriota, Bacteroidota, Pseudoonadota, Thiotrichales, Staphylococcus and Mycobacteriaceae.

After downloading all strain data, in order to facilitate subsequent operations for Cas protein screening process, the genome files in the compressed package are extracted and integrated together.

2.2 Bioinformatics Analysis

2.2.1 Candidate Cas protein screening via CRISPRCasFinder

A CRISPRCasFinder search is then used on the integrated file of all downloaded bacteria genomes to recognize Cas9 protein sequences. CRISPRCasFinder (the updated version of the CRISPRFinder program) is a CRISPR-Cas array detection algorithm by identifying palindromic repeats and checking potential spacers among a sequence. Within which a maximum of 67 Mb query sequence in FASTA format would need to be inputted for the program to detect maximal repeats, of default value 23-55 base pair. Program would then select the DR consensus based on a predetermined score, taking into account the candidate DR's frequency across the entire genome. Candidate CRISPR-Cas segments could be identified then if they are tested to meet fundamental requirements for CRISPR definition (if spacer length is less than repeat length and every spacer is distinct from each other).

Genomes of the previously listed organisms would be screened in search of sequences coding for Cas9 proteins via CRISPRCasFinder, in the process that was described above. After the search is completed, the file of all rawCas protein would be obtained, which will record all Cas protein sequences found in this search, as well as the types and families of proteins they belong to. As is shown in figure1.

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

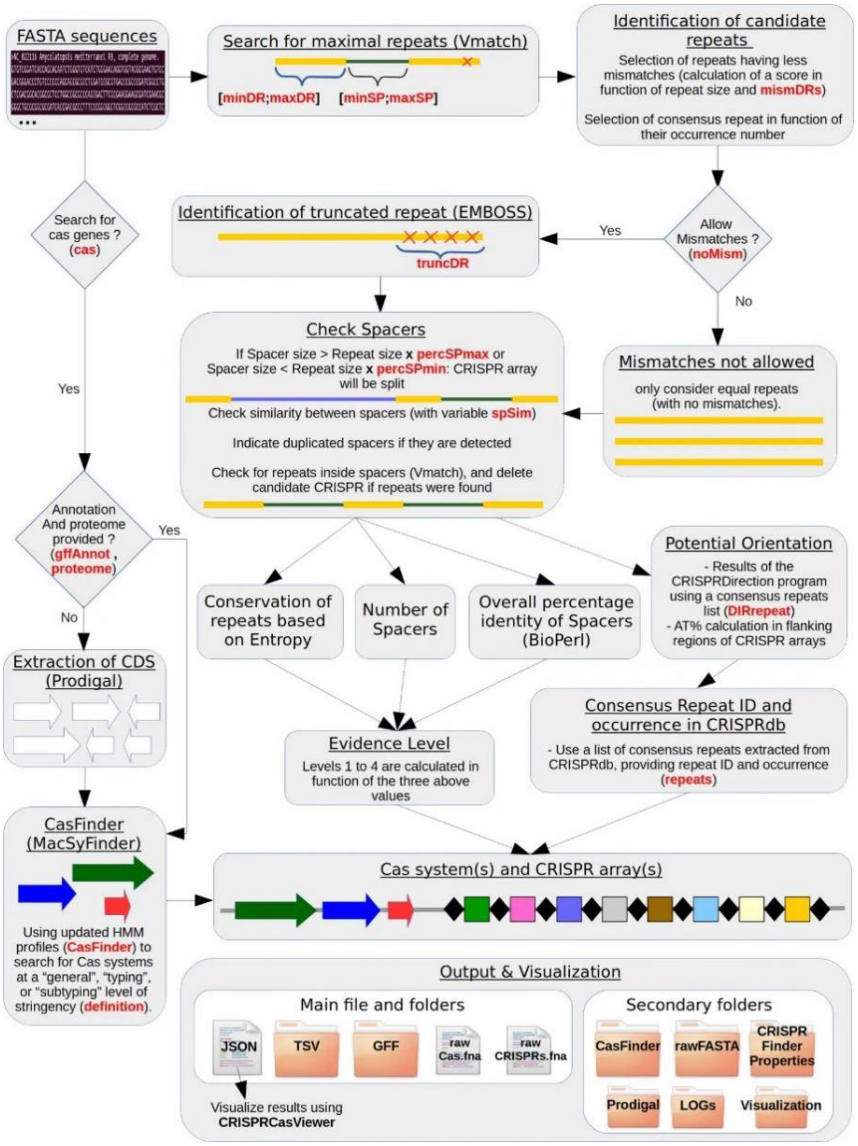

Figure 1: The workflow of the CRISPR-Cas Finder search. The central principle is to search for and identify candidate repeats and spacer fragments.

2.2.2 Removing redundant sequences using SeqKit

Redundant sequences occur when similar sequences are present on the same chromosome, which can introduce biases in data analyses. While the CRISPRCasfinder program identifies candidate Cas protein sequences, it does not filter out redundant sequences, potentially skewing the results.

To address this, SeqKit, a fast toolkit for manipulating FASTA and FASTQ files in Linux, was used to eliminate redundancy in the Cas protein and Cas9 protein sequences. SeqKit efficiently handles tasks such as format conversion, sequence extraction, and quality control on large genomic datasets.

By using SeqKit to remove redundant sequences, a clean dataset is obtained for further analysis and verification of protein functions. The refined output of integrated Cas9 sequences will be utilized in subsequent research and comparisons.

2.2.3 Verification of HNH and RuvC nuclease domain

The HNH and RuvC nuclease domains of Cas9 are crucial for its genome-editing capabilities. The HNH domain cleaves the DNA strand complementary to the guiding RNA, while RuvC cleaves the non-complementary strand. Both cleavage sites are located between the 3rd and 4th bases upstream of the PAM sequence, enabling Cas9's essential double-strand cleavage mechanism. The distinct functions of these domains reflect the precision and versatility of Cas9 proteins.

To verify the cleavage ability of the 44 candidate samples and confirm their classification as Cas

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

proteins, we will examine the presence of HNH and RuvC domains. Positive results would indicate these nuclease domains and suggest the DNA cleavage function of the orthologs.

Protein sequence analysis will be conducted using HMMER, a software for probabilistic searches of protein sequence similarity. HMMER detects homologous sequences by comparing profile hidden Markov models (HMM) with the query sequence, allowing for sensitive searches against large databases, such as Pfam.

Additionally, to ensure the reliability of results, we will use the NCBI Conserved Domain Search (CD-search) plugin. This method employs BLAST and various sequence alignments to identify structural domains in the Cas9 protein sequences and assess the significance of alignments. This dual verification approach aims to enhance the credibility of the structural domain information obtained.

2.2.4 BLASTp comparison of result to known Cas9 sequences

After validating the nuclease domains, the sequences of all candidate Cas9 proteins will be compared to previously discovered Cas9 proteins using the Nucleotide Basic Local Alignment Search Tool (BLASTn). A subsequent verification will involve comparing protein amino acid sequences via the Protein Basic Local Alignment Search Tool (BLASTp).

BLAST tools are categorized by query type and dataset, such as Blastx, which compares translated nucleotides against protein sequences. The algorithm employs a seed-and-extend method, where an 11 bp segment of the query, called a "seed," is matched with the subject sequence. If a match occurs, the T Score (or "hit") is evaluated based on base pair matching. Extension of the seed is performed if the T Score exceeds 18.

In this experiment, the input FASTA-formatted Cas9 nucleotide sequences will be compared to 79 previously identified Cas9 protein sequences from Giedrius Gasiunas and colleagues using BLASTn and BLASTp. Results will be assessed based on each protein-coding nucleotide query's Score and Expectation value.

2.2.5 Phylogenetic representation of new Cas9 orthologs

Orthologs of the Cas9 protein will be selected from the filtered results to construct phylogenetic trees for eight bacterial genomes and archaea, enabling a systematic diversity sampling. The analysis will focus on tree branches to examine relationships between newly discovered and existing Cas9 proteins, with branch lengths reflecting evolutionary changes and providing insights into their origins. Reliability will be assessed through bootstrap testing, considering values over 70% as valuable for sequence origin analysis, while data below this threshold will be excluded but noted for potential errors. Molecular Evolutionary Genetic Analysis (MEGA) software will be used to graph the phylogenetic tree, utilizing the maximum likelihood method. Developed by Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei, MEGA is a widely used tool for analyzing molecular evolutionary data.

2.3 Variational Autoencoder (VAE) model

To make credible predictions of Cas9 protein sequences, this research constructed an auto encoding variational bayes (AEVB), a leaning algorithm composed of an encoder and a decoder. The integration of the VAE model in the process introduces a layer of probabilistic modeling and linear interpolation operations within the latent space to produce results of credible and biologically plausible candidate proteins. As a variant of an Autoencoder, VAE is able to compress the input into a latent space in a low dimensional representation. Usually the data within the latent space would be arranged as normal distribution form, and therefore VAE would be able to generate new data by learning the parameters of the data distribution. After the storage of data, the decoder would then reconstruct the original data from the low dimensional form, while allowing a certain range of assumptions distributions for the generation of new data by the reference of the original input.

During training and utilizing the VAE model, the 79 Cas9 protein sequences will be encoded into a matrix composed of one-hot encoded data as the training set input. During the training process, the encoder pick up the one-hot encoded sequences into the lower-dimensional latent space for data compression. Afterwards, the VAE leverages variational inferences to sort out the best-fitted parameters which minimizes both the construction error and Kullback-Leibler divergence index. The parameters would be modified to fit original input through comparison between the output and input, generating the calculation model to optimize for the scoring of the two items. The details of the VAE is in Figure2

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

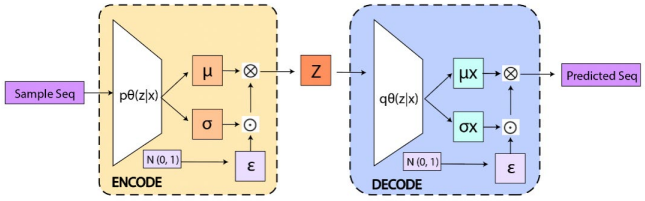

Figure 2: The model struction of VAE.

To offset the inaccuracy in our VAE model that might be caused by this shortage of training set sequences, we managed to incorporate the open-access Variational autoencoder for protein sequences (models/metal16_nostruc) model in the training process of our model. This trained VAE would be able to provide empirical data on protein sequences, providing correction for our modeling process. Due to the limitation in number of existing CRISPR-Cas9 proteins, the exact set of 79 sequences would be inputted into the trained model and to be used as the testing set for output. The expectation is that a new set of 79 CRISPR-Cas9 sequences that is optimized to resemble existing CRISPR-Cas9 sequence would be generated by a VAE model.

Note that the predicted results are not expected to exist in organisms. To ensure the functionality of the predicted candidate proteins, results of Machine training shall be further evaluated on their functionality of gene splicing through other specific testing, such as TracrRNA and PAM sequence verification.

3. Result

3.1 CRISPR-Cas9 protein filter results

We conducted CRISPR sequence filtering on an integrated file of bacterial and archaeal genomes from the JGI-IMG database, which included genera such as Neisseria, Staphylococcus, Mycobacteriaceae, Thiotrichales, Acidobacteriota, Bacteroidota, Pseudomonadota, and general archaea. This analysis was performed using CRISPRCasFinder software, with bacterial groups organized from smallest (genus) to largest (domain) classification levels. Out of 25,495 bacterial and archaeal strains, we identified 3,487 Cas protein-encoding gene segments, including 44 confirmed candidate Cas9 proteins (Appendix Table 1).

Neisseria showed the highest Cas9 content in our screening of 712 bacterial strains, from which we extracted 710 files. This revealed 158 non-redundant CRISPR inverted repeat sequences and 93 associated Cas protein family genes, with 14 confirmed as Cas9. Type I CRISPR-Cas systems were the most abundant, as evidenced by the high prevalence of Cas3 proteins, while Type II systems, which utilize Cas9 as the primary effector, ranked as the second most common (see Figure 3).

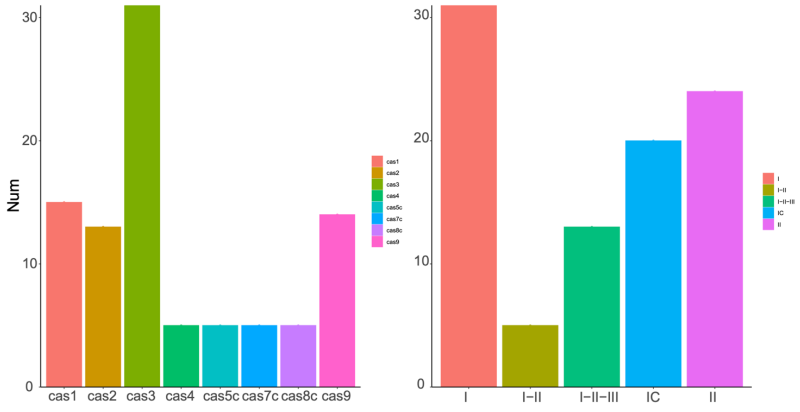

Figure 3: Classification of Cas proteins obtained from the Neisseria.

While Streptococcus yielded a higher number of Cas9 protein sequences from a larger dataset. Initially, 3,606 bacterial strains were screened, resulting in 3,605 genomic files extracted. This analysis identified 109 CRISPR inverted repeat sequences and 218 Cas proteins, with 21 belonging to the Cas9 family. Cas3 proteins were the most prevalent, outnumbering the Streptococcus Cas9 proteins.

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

Consequently, Type I CRISPR-Cas systems were the most common, closely followed by Type II systems (see Figure 4).

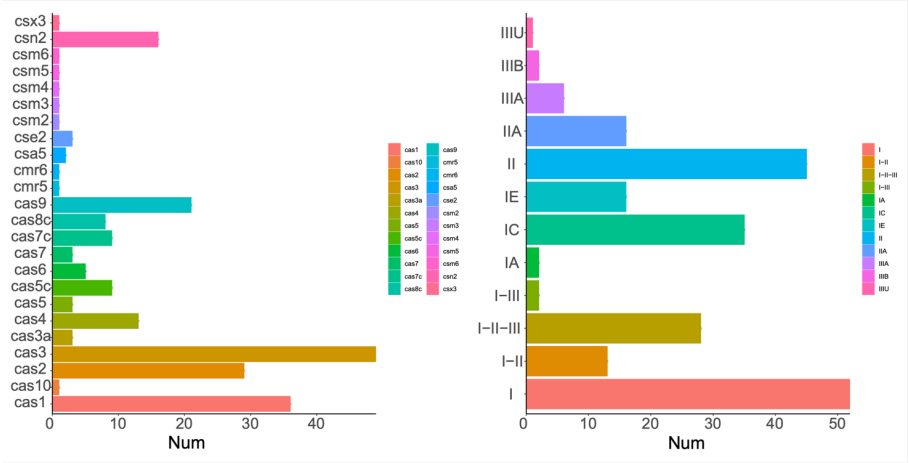

Figure 4: Classification of Cas proteins obtained from the Streptococcus.

Among the 4,215 utilized and successfully extracted Bacteroidota genome files, 50 exhibited non-redundant palindromic repeat sequences, leading to the identification of 54 Cas protein segments. Surprisingly, only one candidate protein belonged to the Cas9 family. Cas3 proteins were the most common among the screened Bacteroidota Cas protein sequences, while the proportion of Cas9 proteins was relatively low compared to other Cas proteins. Consequently, Type I Cas systems accounted for the majority of candidate systems, with Type II systems making up only about one-fifth of the Type I systems (see Figure 5).

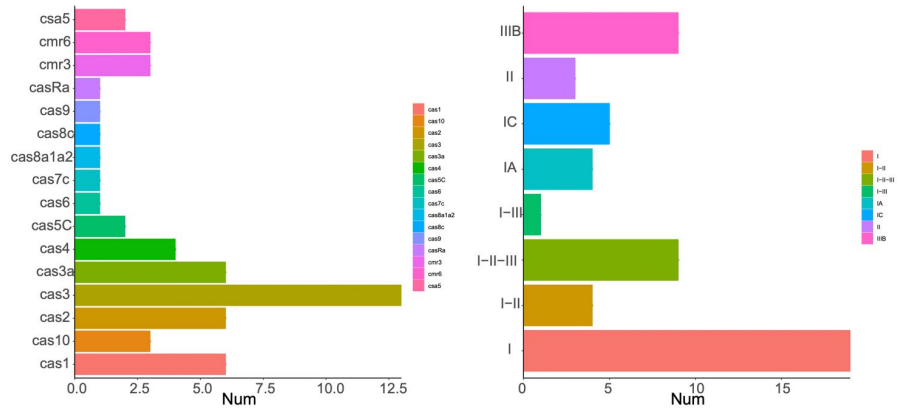

Figure 5: Classification of Cas proteins obtained from the Streptococcus.

A total of 5,780 files were successfully extracted from 39,251 genomes in the Pseudomonadota genome bank for screening. We identified 2,111 sequences with palindromic repeats, totaling 2,777 candidate Cas protein systems. Seven non-redundant candidate Cas9 proteins were identified in the selected set. Pseudomonades exhibited a diverse range of 42 CRISPR-Cas protein systems, significantly more than other bacterial phyla. Cas3 proteins dominated the candidate Cas proteins, while Cas9 proteins accounted for a small percentage. Type I CRISPR-Cas systems were the primary type, followed by Type I-E systems, which had about half the number of Type I (see Figure 6).

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

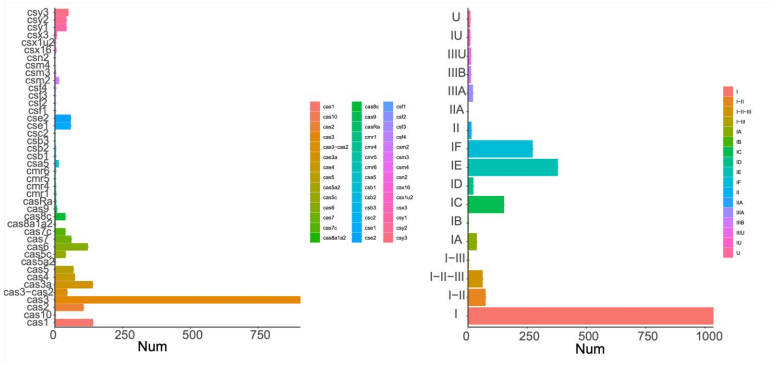

Figure 6: Classification of Cas proteins obtained from the Pseudomonadota.

A total of 6,078 sequences from the Staphylococcus genome were selected for screening, with 6,076 successfully extracted. Among these, 148 sequences contained palindromic repeat segments, leading to the identification of 51 Cas proteins. Only one candidate Cas9 protein was found in the Staphylococcus genome database. The CRISPR-Cas3 system was notably abundant compared to other CRISPR-Cas systems, while the CRISPR-Cas9 system was rare among the candidates. Type I CRISPR-Cas systems were the most common, followed by Type III-A (see Figure 7).

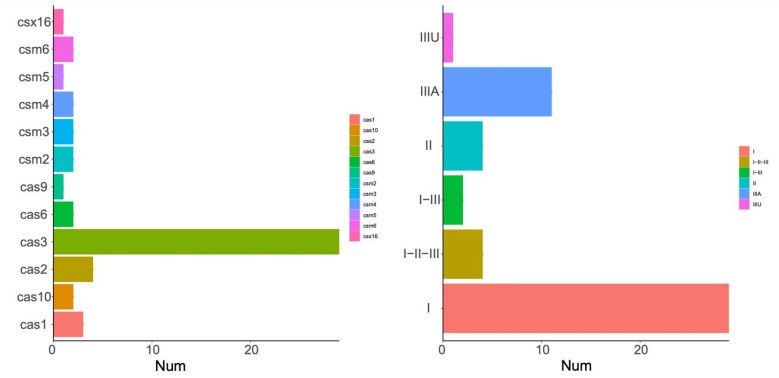

Figure 7: Classification of Cas proteins obtained from the Staphylococcus.

Among the eight bacterial genomes and archaea sequences, candidate CRISPR-Cas9 proteins were successfully identified in five genomes. However, the filtering for four categories (Acidobacteriota, Thiotrichales, Mycobacteriaceae, and Archaea) revealed Cas protein structures but no specific CRISPR-Cas9 candidates.

Acidobacteriota: From 73 extracted genomes, 21 CRISPR palindromic repeats and 40 non-repetitive Cas protein segments were found. Thiotrichales: In 366 extracted genomes, 29 repeats and 10 Cas proteins were identified. Mycobacteriaceae: Among 4,744 extracted genomes, 215 CRISPR repeats and 84 Cas proteins were found. Archaea: From 1,973 extracted genomes, 49 palindromic repeats and 160 Cas protein segments were identified.

3.2 Structural domain (HNH, RuvC) comparison

Domain annotation of the candidate Cas9 proteins was conducted using methods such as the NCBI Conserved Domain Search (CD-search) plugin and retrieval from the Protein Families (Pfam) database. This approach allowed for mutual verification of the structural content of the proteins through these two distinct methods. Several filtered results displayed typical structural domains of Cas9 proteins, as evidenced by the comparison of protein structural domains from the screened clean outcomes.

The results from the Neisseria CD-search (Appendix table 2) and pfam annotation table (Appendix table 3) indicated that the detected sequences were categorized based on specific and significant matches, with 10 out of 14 candidate outcomes showing the presence of essential nuclease domains. And The results from the Streptococcus CD-search(Appendix table 4)and pfam annotation table(Appendix table 5)showed that among all 21 candidates, 9 filtered Cas9 proteins were found to possess specific hits and significant matches, indicating their presence of the target nuclease domain. The results from the Streptococcus CD-search (Appendix table 6) and pfam annotation table(Appendix table 7)indicated that

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

among all 21 candidates, 9 filtered Cas9 proteins possessed specific hits and significant matches, confirming their presence of the target nuclease domain. And the results from the Pseudomonadota CD-search(Appendix table 8)and pfam annotation table(Appendix table 9)indicated that among the 9 screened Cas9 sequences, 3 contained specific matches to the HNH and RuvC nuclease domains.

3.3 BLAST comparison

BLAST comparison results were evaluated based on Score and E-value, where smaller E-values and larger Scores indicate higher homology. To identify new Cas9 proteins, target samples must differ from the 79 existing proteins; those with an E-value of 0 or a Score over 70 were discarded due to homology.

A total of 44 protein sequences underwent filtering, with 21 found to differ from the existing Cas9 proteins. The candidates with low homology are listed below:Neisseria: 3 of 14 queries showed little homology (cas9_TypeII_1 36284, 39577; cas9_TypeII_1 27639, 30884; cas9_TypeII_1 721, 2826). Streptococcus: 12 of 21 queries differed (e.g., cas9_TypeII_1 1456318, 1459686; cas9_TypeII_1 1276484, 1280611; etc.). Staphylococcus: 1 of 1 query differed (cas9_TypeII_1 144826, 146253). Pseudomonadota: 4 of 7 queries differed (e.g., cas9_TypeII_1 4143336, 4146401; cas9_TypeII_1 36284, 39577; etc.). Bacteroidota: 1 of 1 query differed (cas9_TypeII_1 81827, 86086).

3.4 VAE Model

We next evaluated the learning outcome of our Deep learning model. After training, the model is adjusted to well form. The coding for the CRISPR-Cas9 VAE generator is adjusted based on the VAE model protein-vae, which utilizes the VAE algorithm to generate similar amino acid sequences based on known amino acid sequences. We obtained 79 prediction sequences after the input of the newly screened out results from the bacterial genome database, which is displayed below. Again, future research should be conducted focusing on characterizing these predicted sequences in wet lab experiments to confirm their properties and assess their practical utility in genome editing.

3.5 Phylogenetic tree

The vertical phylogenetic tree (see Figure 8) of 44 can0didate Cas9 proteins revealed two main clades: one predominantly consisting of Streptococcus, Bacteroidota, Pseudomonadota, and Staphylococcus, and the other comprising Neisseria, with a bootstrap value of 70. The Streptococcus clade included a Pseudomonadota sister strain (11999, 15055) with a posterior probability of 91.

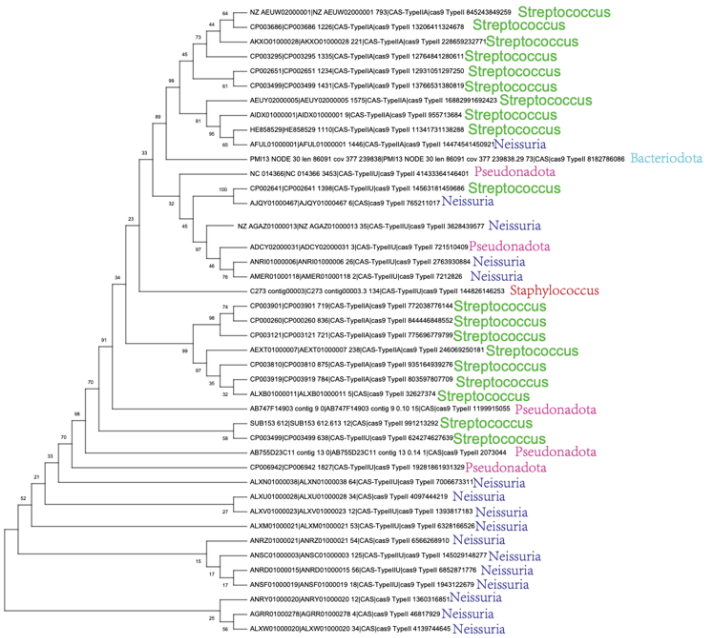

Figure 8: The phylogenetic tree illustrates the homology between the screened Cas9 protein sequences.

The Neisseria cluster showed a progressive differentiation pattern, forming distinct clades, with high

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

distinction among sequences like 145029148277 and 6328166526. In contrast, Streptococcus sequences displayed greater similarity, as indicated by high bootstrap values among individuals like 844446 and 848552. Identical Cas9 sequences, such as Streptococcus 14563181459686 and Neisseria 765211017, suggest origins from different bacterial phyla, with a 100% bootstrap match. Most Cas9 types diverged into three distinct clusters, indicating potential unique properties and functionalities.

At the tree's top, a cluster of new proteins with shorter branches suggests high sequence similarity and common functional characteristics. The observed divergence implies that these new Cas9 proteins may address limitations in current proteins, such as target specificity and editing efficiency, offering novel tools for genome editing and genetic engineering.

4. Discussion

The results of this study have significant implications for the field of gene editing and therapeutic research: The identification of 44 new candidate Cas9 proteins from various bacterial genomes, including Neisseria, Streptococcus, and Pseudomonadota, represents a substantial advancement in our understanding of CRISPR-Cas systems. These newly discovered Cas9 proteins exhibit unique properties that could potentially address some of the limitations associated with the currently used SpyCas9 protein.

Using the JGI-IMG database, this experiment identified Cas9 proteins in various bacterial genomes, expanding the known range of Cas proteins. Future studies could investigate whether the newly discovered 44 proteins have different applications compared to the original 79, such as differing PAM sequences or enhanced cleavage specificity to reduce off-target effects. Validation of the HNH and RuvC nuclease domains confirmed that the screened proteins maintain traditional Cas9 cleavage functions. However, some sequences lacking these domains may still possess cleavage abilities, necessitating further verification of alternative domains or optimization through protein engineering for improved gene editing.

The phylogenetic analysis revealed evolutionary differences between the new and existing proteins, indicating functional distinctions. If subsequent wet lab experiments and structural validations show these 44 proteins perform better than SpyCas9 under specific conditions, the next step would involve applying them in research and clinical settings.

Unfortunately, no Cas9 proteins were found in archaea or some bacterial databases, limiting the study's scope to a subset of bacterial genomes and potentially missing valuable Cas9 variants. Future experiments should broaden the search to include more bacterial strains, increasing the chances of discovering Cas9 variants with novel functionalities or enhanced performance. Exploring broader soil and marine metagenomic data may reveal additional Cas9 protein candidates with unique properties.

For the 44 new protein sequences, further verification could involve comparing their tracrRNA and crRNA with SpyCas9. The uniqueness and limitations of CRISPR-Cas9 hinge on specificity for both binding RNA and recognition sequences. Trans-activating CRISPR RNA (tracrRNA) pairs with CRISPR RNAs (crRNA) to form the single guiding RNA (sgRNA) essential for target sequence identification. TracrRNA, found within Cas9 proteins, serves as a common identification feature. Additionally, the Protospacer Adjacent Motif (PAM) sequence—short sequences (2-6 base pairs) on the non-complementary strand of target DNA—helps Cas9 determine cleavage points and indicates the origin and type of Cas9 based on bacterial species.

To ensure the specificity of the newly discovered Cas9 proteins, future research should focus on comparing PAM and tracrRNA sequences to those of SpyCas9 through wet lab procedures. Any differences observed would highlight the novelty of the new proteins and suggest broader applications for their use.

On the other hand, for the 79 generated results of the deep learning model VAE, this experiment used protein models and sample from existing 79 proteins to simulate possible Cas9 protein sequences. These sequences need to be tested for their nuclease domains and for more verification methods in the future to confirm whether they have functions of the Cas protein family. Due to the scarcity of existing Cas9 proteins, the VAE model wasn’t able to generate enough simulated protein sequences while ensuring accuracy. Future models could enhance the accuracy of sequence simulation and generate more simulated protein sequences by conducting analysis on the newly discovered 44 Cas9 proteins.

ISSN 2616-5910 Vol. 5, Issue 1: 135-144, DOI: 10.25236/AJALS.2024.050118

5. Conclusion

In recent years, the CRISPR Cas system, particularly the Type II CRISPR Cas9 system, has advanced significantly in biomedical research due to its precision in gene editing. This system aids in understanding gene function and disease mechanisms and shows promise for treating tumors and multi-gene diseases. This study identified 44 non-redundant candidate Cas9 proteins from various bacterial genomes using the JGI-IMG database and SeqKit, providing a valuable resource for developing precise gene editing tools.

HMMER and NCBI's CD-search confirmed the presence of essential HNH and RuvC nuclease domains in these proteins, ensuring their functionality. Phylogenetic analysis and domain verification suggest these new Cas9 proteins could address limitations of the widely used SpyCas9. Additionally, 79 Cas9 protein sequences were predicted using a trained VAE deep learning model. The integration of bioinformatics tools and deep learning has proven effective for discovering and predicting new Cas9 proteins, though further experimental validation is needed to confirm their functionality and specificity.

In conclusion, these findings enhance our understanding of CRISPR-Cas systems and contribute to developing innovative gene editing tools that could significantly impact genetic disease treatment and biomedical research advancement.

References

[1] Gupta, Darshana, et al. "CRISPR-Cas9 system: A new-fangled dawn in gene editing." Life sciences 232 (2019): 116636.
[2] Ma, Yuanwu et al. "Genome modification by CRISPR/Cas9." The FEBS journal vol. 281, 23 (2014): 5186-93. doi:10.1111/febs.13110
[3] Frangoul, Haydar, et al. "CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia." New England Journal of Medicine 384.3 (2021): 252-260.
[4] Maxwell, Kristina G., et al. "Gene-edited human stem cell–derived β cells from a patient with monogenic diabetes reverse preexisting diabetes in mice." Science translational medicine 12.540 (2020): eaax9106.
[5] Zhao, Huan, et al. "In vivo AAV-CRISPR/Cas9–mediated gene editing ameliorates atherosclerosis in familial hypercholesterolemia." Circulation 141.1 (2020): 67-79.
[6] Razeghian, Ehsan et al. "A deep insight into CRISPR/Cas9 application in CAR-T cell-based tumor immunotherapies." Stem cell research & therapy vol. 12, 1 428. 28 Jul. 2021, doi:10.1186/s13287-021-02510-7
[7] Carroll, Kelli J et al. "A mouse model for adult cardiac-specific gene deletion with CRISPR/Cas9." Proceedings of the National Academy of Sciences of the United States of America vol. 113, 2 (2016): 338-43. doi:10.1073/pnas.1523918113
[8] György, Bence et al. "CRISPR/Cas9 Mediated Disruption of the Swedish APP Allele as a Therapeutic Approach for Early-Onset Alzheimer's Disease." Molecular therapy. Nucleic acids vol. 11 (2018): 429-440. doi:10.1016/j.omtn.2018.03.007
[9] Bhushan, Kul, Anirudha Chattopadhyay, and Dharmendra Pratap. "The evolution of CRISPR/Cas9 and their cousins: hope or hype?" Biotechnology letters 40 (2018): 465-477.
[10] Zhang, Weiwei, et al. "In-depth assessment of the PAM compatibility and editing activities of Cas9 variants." Nucleic Acids Research 49.15 (2021): 8785-8795.
[11] Markowitz, Victor M et al. "IMG: the Integrated Microbial Genomes database and comparative analysis system." Nucleic acids research vol. 40, Database issue (2012): D115-22. doi:10.1093/ nar/ gkr 1044