Structured Learning on Bayesian Networks for Guide Efficiency Prediction for CRISPR-CAS13 System

Dr. Saleha Raza
Associate Professor, CS
Habib University
Karachi, Pakistan
saleha.raza@sse.habib.edu.pk

Syed Muhammad Ali Naqvi
Computer Science
Habib University
Karachi, Pakistan
sn07590@st.habib.edu.pk

Musab Kasbati
Computer Science
Habib University
Karachi, Pakistan
mk07811@st.habib.edu.pk

Ali Muhammad Asad
Computer Science
Habib University
Karachi, Pakistan
aa07190@st.habib.edu.pk

I. Introduction

CRISPR-Cas based drugs are revolutionizing the field of genetic medicine. These drugs work by using a specialized gene-editing technology originally discovered in bacteria. This system helps bacteria cut out harmful viral DNA from their own genomes to defend against infections. Scientists have adapted this mechanism for therapeutic purposes in humans, particularly for treating viral infections. Viruses, like HIV, hepatitis B, or Dengue, integrate their genetic material into the DNA of human cells. CRISPR-Cas technology is designed to recognize and cut out these viral sequences from infected cells. Until the development of CRISPR-Cas technology, introducing site specific changes into a DNA sequence was a challenging task, thus limiting the potential of genetic research.

CRISPR systems, typically use two main components: an enzyme that acts like molecular scissors, and a guide RNA (gRNA) that directs the enzyme to the viral DNA/RNA sequence. Once the guide RNA leads the enzyme to the viral DNA/RNA within human cells, it makes precise cuts in the DNA. This action can disrupt the viral genome, preventing the virus from replicating and spreading further.

There are several types of CRISPR system typically differentiated based on their mechanism and structure. The most popular of which is CRISPR CAS 9 system which is used to cleave DNA. A recent variant, CRISPR-Cas13, targets RNA molecules rather than DNA, using programmable guide RNAs (gRNAs) to selectively bind and cleave target RNA sequences with high precision and effectiveness [1]. Cas13b can directly target the viral RNA, stopping the virus from producing proteins necessary for replication. This is especially useful for rapidly evolving RNA viruses where traditional treatments struggle. This makes CRISPR-Cas13 a powerful tool for antiviral treatments, where targeting viral RNA can inhibit infections.

The design of the guide crRNA (CRISPR RNA) is extremely crucial since it determines where the CRISPR-Cas molecule will make the cuts. The guide needs to be very precise to make sure the infected part of the DNA is cut. A poorly designed guide crRNA can lead to unwanted changes in the DNA of the cells. The efficacy of a gRNA is evaluated on the basis of two main factors: guide efficiency and guide specificity. Guide efficiency relates to the intensity of activity of a gRNA on target sites, while guide specificity, focuses on the minimization of effect when the gRNA is off-target. However, determining the features which positively or negatively affect the efficiency and specificity of gRNAs remains a challenge. Currently, such features are largely determined through manual testing and empirical methods, requiring extensive lab work. While machine learning/deep learning based computational models have been developed to predict the efficiency and specificity of potential gRNA sequences for Cas 9 and Cas 13 systems, they lack in terms of providing insights into the interpretability and explanation of the features of the sequence which directly affect the efficiency of the gRNA.

II. Related Works

CRISPR-Cas Technology offers a versatile and powerful tool for genome editing, with wide applications in fields such as functional genomics, immunotherapy, synthetic lethality, drug resistance, matastasis, genome regulation chromatic assessiblity and RNA-targeting [2].

We came across a study by Han, in which an Elastic-Net model was trained to predict the efficacy of the testing sequence data [3]. They designed two experiments, inter species and within library. In the inter species experiment, DNA sequences belonging to the human ribosomal gene and human non-ribosomal gene were used for training purposes, and sequences of mESC genes from mice were used for testing purposes. In the within libary experiment, the training was done on human ribosomal genes, while testing was done on human non-ribosomal genes. They also carefully selected 28 features from 160 features for a 40-nt DNA sequence, as they showed statistical significance and biologically reasonable correlations to the efficiency measurement of the genes they trained and tested their data on [3]. Their implementation of Elastic Net is LASSO, to select a smaller number of features from a larger amount of candidates. Their model achieves significant results in both positive and negative selection conditions.

Another study by Doench et al. proposed a logistic regression classifier model to discriminate the highest activity sgRNA (single guide RNA) [4], to later discover that additional features such as position-independent nucleotide counts and the locations of the sgRNA target sites can impact the classification results for the better [5]. In their study, they generated specific features using the L1-regularized linear support vector machine, such as 39 single nucleotide features, 30 dinucleotide features, and 2 GC counts. Then they trained a logistic regression classifier model on 8 of the genes, and tested it on 1 gene. In total 9 logistic regression classifiers were implemented 9 genes, called the "Replication-Doench". Their parameters weren't provided in the literature.

A more general overview of the various efficacy prediction tools for CRISPR-Cas9 gRNAs was provided by Konstantakos et al. [6]. In general, the focus has been mainly on three types of tools: hypothesis-based scoring, machine-learning models, and deep-learning models. Hypothesis-based tools, such as E-CRISP, score guide RNAs based on sequence features that have been hypothesised to generally lead to more effective gRNAs. On the other hand, machine-learning models, like Azimuth 2.0, and deep-learning based models, such as CRISPRLearner and DeepCRISPR, take a data-driven approach searching for correlations in data between gRNA sequences and experimentally evaluated efficacy. They then try to use these correlations to generalise to untested gRNAs for untested target RNAs. However, evaluating these models have shown that deep-learning based models tend to generalise poorly to datasets other than those they are trained upon, and the gap between hypothesis-based models and deep-learning based models in terms of accuracy is not as vast as one might initially expect, with E-CRISP generalising better across datasets. Moreover, these deep-learning models tend to lack the interpretability that is often desired by domain specialist who wish to expand upon the general understanding of how CRISPR works. Yet, having a plethora of predictive tools available has had the benefit of reinforcing hypothesis on gRNA behaviour when multiple tools end up giving weight to a common set of sequence features.

Previous work done by Yi Yan employs a Bayesian Network to model the relationship between sequence features of the target DNA, and the efficacy of the CRISPR-Cas9 system [7]. Their work starts by replicating the results of 2 studies mentioned above [3] [4], [5]. They then explain why Naive Bayes (a class of Bayesian Networks) works better as a generative model than Logistic Regression, adopting Bayesian network structure learning. For the Han Data, their Naive Bayes model had 28 feature nodes, and 1 label node, as Han chose 28 features. For the Doench data, they selected 27 features as 1 of the features listed by Han was not available in the Doench data. Their Naive Bayes classifier gave similar AUC scores compared to the replication of Han and Doench [7].

For discovering relationships among feature nodes, they adopted Bayesian structure learning. Structured learning involves first learning the topology of the graph, followed by parameter learning - conditional probabilities, then by inference to predict the efficiency of the DNA sequences in the study. They used Hill Climbing algorithm through "pgmpy" package provided by Python for a score based structure learning where the algorithm tries to find a local maxima of the score function. They used the BDeu score function [7]. They modify the Hill climbing algorithm to not delete the Naive Bayes edges, and restricted the algorithm to not repeat the previous 50 actions. They named this approach "constrained structure learning". When they allow the deletion of the nodes, they call it non-constrained structure learning. They then move onto solving the false conditional independence of nucleotides assumption by changing the dummy encoding to k-mer encodings, as by dummy encoding they assume that all nodes are independent of each other given the label node, however, that is not the case in truth, as for a fixed DNA sequence, the nucleotides are ACGT, and while a position is one of the nucleotide, it cannot be the other. Finally they use D-Seperation to understand the mechanics of CRISPR-Cas9 to find the location of the active site of the Cas9, and the location of scissile bonds was consistent with their D-Seperation findings [7].

CRISPR-Cas13 is a more recent technology that targets RNA instead of DNA, offering high sensitivity and specificity for the detection of microorganisms through programmable RNA guides [1], [8]. Due to the novelty of this technology, current methods for determining gRNA efficiency rely heavily on empirical testing in lab settings. However, there exists some computational models such as CASowary [9] which is a machine-learning model which generates a shortlist of candidate sgRNAs for a given target RNA. Another is TIGER [10], which is a deep-learning model that was trained on many sample gRNAs, some with intentionally introduced sequence errors, in order to better infer the impact of sequence structure and target RNA context upon gRNA efficacy.

Yet, no computational method has been explored for the interpretability and explainability of features of gRNA affecting its efficacy in Cas13 systems.

REFERENCES

[1] D. B. T. Cox, J. S. Gootenberg, O. O. Abudayyeh, B. Franklin, M. J. Kellner, J. Joung, and F. Zhang, "Rna editing with crispr-cas13," *Science*, vol. 358, no. 6366, pp. 1019–1027, 2017.

[2] M. Colic and T. Hart, "Common computational tools for analyzing crispr screens," *Emerging Topics in Life Sciences*, vol. 5, 12 2021.

[3] H. Xu, T. Xiao, C.-H. Chen, W. Li, C. A. Meyer, Q. Wu, D. Wu, L. Cong, F. Zhang, J. S. Liu, M. Brown, and X. S. Liu, "Sequence determinants of improved crispr sgrna design," *Genome Research*, vol. 25, no. 8, 2015.

[4] J. Doench, E. Hartenian, D. Graham, Z. Tothova, M. Hegde, I. Smith, M. Sullender, B. Ebert, R. Xavier, and D. Root, "Rational design of highly active sgrnas for crispr-cas9-mediated gene inactivation," *Nature biotechnology*, vol. 32, 09 2014.

[5] J. Doench, N. Fusi, M. Sullender, M. Hegde, E. Vaimberg, K. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, W. Virgin, J. Listgarten, and D. Root, "Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9," *Nature biotechnology*, vol. 34, 01 2016.

[6] V. Konstantakos, A. Nentidis, A. Krithara, and G. Paliouras, "CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning," *Nucleic Acids Research*, vol. 50, pp. 3616–3637, 03 2022.

[7] Y. Yan, "Efficiency prediction and mechanism discovery for the crispr-cas9 system," 05 2018.

[8] Z. Huang, J. Fang, M. Zhou, Z. Gong, and T. Xiang, "Crispr-cas13: A new technology for the rapid detection of pathogenic microorganisms," *Frontiers in Microbiology*, vol. 13, 2022.

[9] A. Krohannon, M. Srivastava, S. Rauch, R. Srivastava, B. C. Dickinson, and S. C. Janga, "Casowary: Crispr-cas13 guide rna predictor for transcript depletion," *BMC Genomics*, vol. 23, no. 1, p. 172, 2022.

[10] H.-H. Wessels, A. Stirn, A. Méndez-Mancilla, E. J. Kim, S. K. Hart, D. A. Knowles, and N. E. Sanjana, "Prediction of on-target and off-target activity of crispr–cas13d guide rnas using deep learning," *Nature Biotechnology*, vol. 42, no. 4, pp. 628–637, 2024.