Minimizing Off-Target Effects of CRISPR-Cas9 with Optimized sgRNA:Evalution of Efficiency and Specificity in the Tumor Protein p53 (TP53) Region

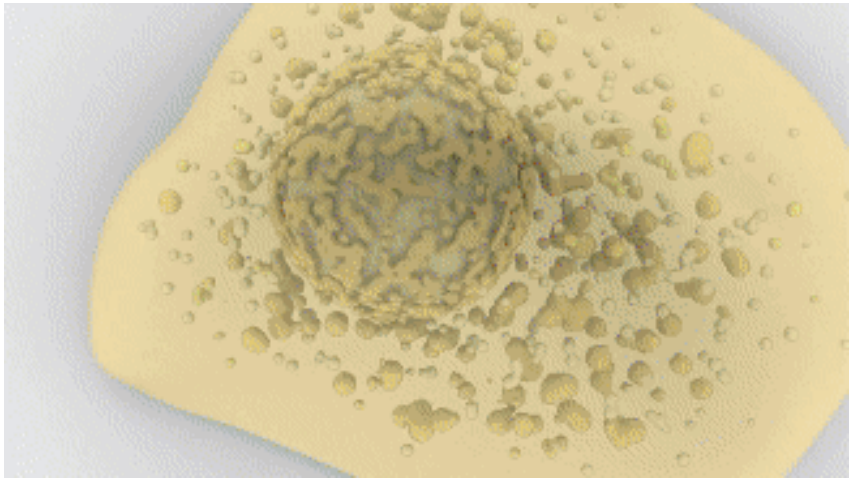
Ali Mertcan Köse Msc.[1],Monia Ranalli Ph.D.[2]
[1]Istanbul Ticaret University-Department of Computer Programming
[2]Sapienza University-Department of Statistics
alimertcankose@gmail.com, monia.ranalli@uniroma1.it

21 August 2024

SAPIENZA
UNIVERSITÀ DI ROMA

GenE-HumDi

Outline



- Introduction
- Material and Methods
- Application
- Discussion
- References

Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is one of the most exciting developments in the field of genome editing technology in molecular biology in recent times (Fusi et al., 2015).

Introduction

What is the CRISPR Application?

- CRISPR is a bacterial and archaeal immune system that combats viral structures within the DNA sequence.
- Cas proteins and guide RNA make modifications in target genetic sequences.
- Potential applications: Treatment of genetic diseases, strengthening of the immune system.

Introduction

- Homology Directed Repair (HDR): Genetic modification in the targeted DNA sequence.
- Non-Homology End Joining (NHEJ): Repair of DNA sequences.

Introduction

The genetic procedures conducted for the detection of nucleotides in the CRISPR locus are carried out based on the prediction of two regions: Off-Target and On-Target. According to the predicted results, Off-Target/On-Target mutations are expected to occur in a specific manner in the size and repeat number of the CRISPR locus (Zischewski et al., 2017).

Introduction

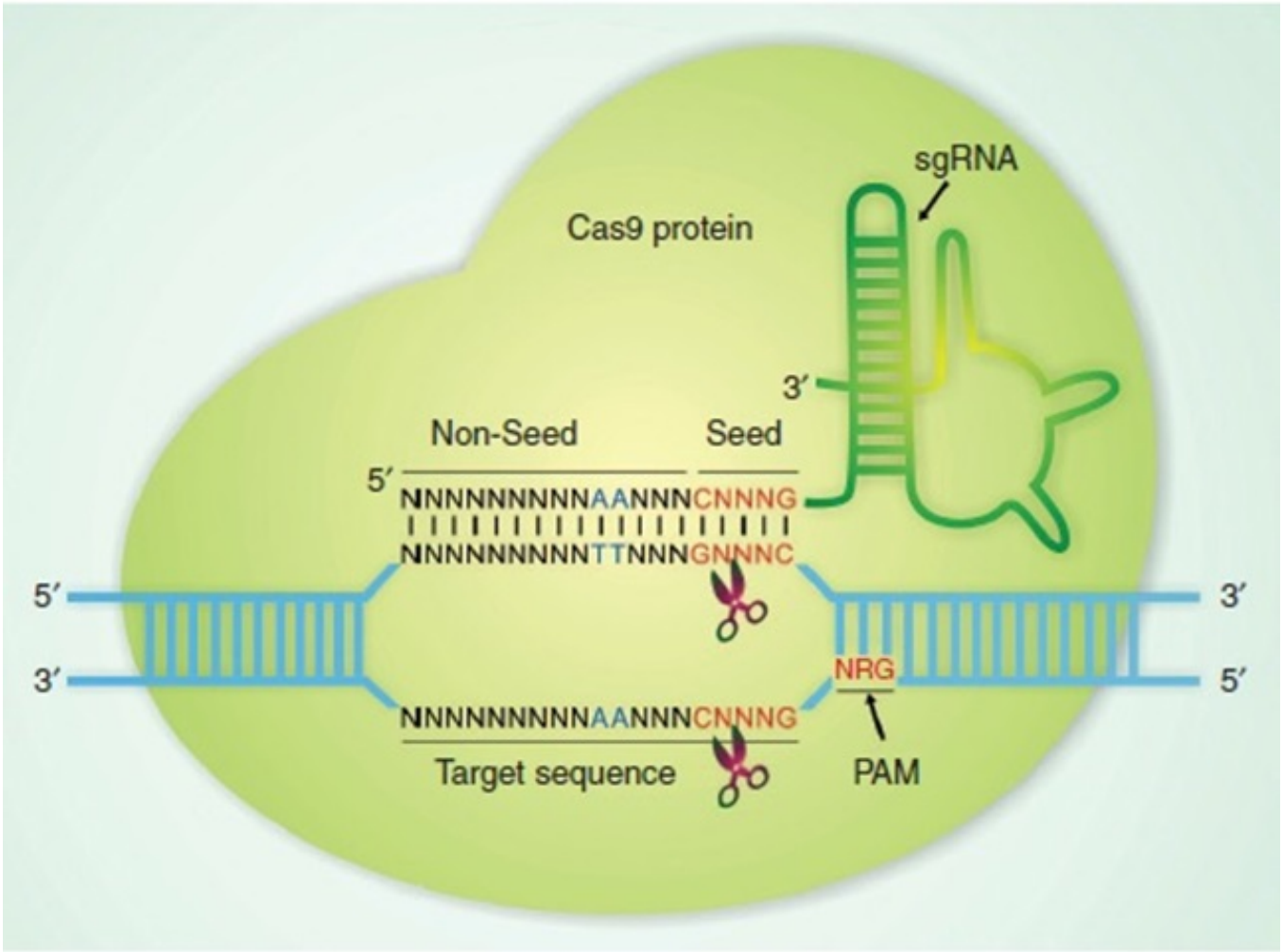


Figure 1: CRISPR Cas System.

Introduction

Off-Target ve On-Target

- To potentially determine Off-Target scores, there are various prediction methods based on mismatch positions, such as CFD, MIT, MIT website, Cropit, and CCTop Score.
- According to Haeussler et.al. (2016), the CFD score has the highest accuracy rate at 91%. On the other hand, the MIT score shows an accuracy of 87%, the MIT website 73%, the Cropit score 81%, and the CCTop Score 77%.

Introduction

Goal

In this context, instead of using only two-level groups such as Off-Target and On-Target, determining two or more class levels through latent class analysis helps in more accurately interpreting and classifying effectiveness scores. By applying machine learning and deep learning approaches alongside latent class analysis, the positions of these mismatches can be predicted. This will pave the way for estimating more precise multi-class models, rather than just two-level classification models.

Introduction

Thus, by using multiple Off-Target classes, it will be possible to systematically test the mismatch effect and ensure that the minimum nucleotide changes are made at each position.

Latent Class Analysis

The main goal of Latent Class Analysis (LCA) is used to identify subgroups in range of substantive areas, including differential diagnosis among disorders. LCA is an analytic technique that has become increasingly popular among researchers. LCA is a technique for analysing relationships in categorical data that is relationships among variables scored at either the nominal or ordinal level of measurement.

Latent Class Analysis

LCA is similar to cluster analysis with respect to theoretical background. LCA has various advantages. It is model based or probabilistic that can be reproduced by using an independent sample which cannot be performed with cluster analysis. Besides, LCA produces statistical fit indices which gives the opportunity to evaluate the model fit and to determine the number of classes.

		Manifest variables y_i	
		Continuous	**Categorical**
Latent variables x_j	**Continuous**	Factor analysis	Latent trait analysis
	Categorical	Latent profile analysis	Latent class analysis

Adapted from Bartholomew and Knott (1999, p.3)

Figure 2: Type of Latent Variable Models.

Machine Learning

The process of extracting information from data is called “Machine Learning.” Machine Learning is a field at the intersection of Statistics, Artificial Intelligence, and Computer Science. Additionally, Machine Learning can be referred to as predictive analytics or statistical learning. The primary aim of Machine Learning is to analyze data using statistical methods based on the obtained data. It consists of a set of methods that can automatically detect patterns in the data and then use the discovered patterns to predict future data.

Materials and Methods

Data Sources

In the application, analyses were performed using the Homo sapiens tumor protein p53 (TP53) dataset, RefSeqGene (LRG_321) on chromosome 17 dataset link. The dataset consists of 9 variables: gRNA, off-target sequence, mismatch position, mismatch count, MIT off-target score, chromosome, start-end point, strand, and location region. In this dataset, there are a total of 23 endonucleases and 947 possible location regions for gRNA and target sequences.

Data Pre-processing and Encoding

The target sequence on DNA and gRNA consists of 23 endonucleases. Considering that the combination of two bases in the same region forms a variable, the DNA sequence is composed of four bases: A/C/G/T. From these four bases, there are 4x4 = 16 possible combinations. These possible matches and mismatches between the target sequence on DNA and gRNA are expressed as One-hot encoding.

Multi-Categorical Coding

Table 1: Multi Categorical Coding

gRNA	Target Squence	Matching
A	A	1
C	A	2
G	A	3
T	A	4
C	C	1
A	C	5
G	C	6
T	C	7
G	G	1
A	G	8
C	G	9
T	G	10
T	T	1
A	T	11
C	T	12
G	T	13

One-Hot Coding

	G	G	T	G	A	C	T	A	A	C	G	T	T	C	A	G	T	C	T	A	G	G
A	0	0	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0
T	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0
C	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
G	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	1

Figure 3: One-Hot Coding.

Materials and Methods

Table 2: Model Features

Features	Target
V_1	
V_2	
V_3	
V_4	
V_5	
V_6	
V_7	
V_8	
V_9	
V_10	
V_11	
V_12	
V_13	
V_14	
V_15	
V_16	
V_17	
V_18	
V_19	
V_20	
V_21	Class
V_22	levels
V_23	
mismatchCount	
chrom	
strand	

Application

After preprocessing and encoding the data, the analysis was performed using Python, R, and Mplus software. The latent class analysis method was used to establish a specific level based on the number of mismatched endonucleases in the CRISPR genome. Latent class analysis was applied separately to binary and multi-coded datasets.

Application

Table 3: Latent Class Analysis Results

Class	Log-Likelihood	AIC	BIC	SSABIC	CAIC	AWE	LMR Test	Entropy
2	-21753.8	44405.6	46584.7	45158.7	47033.7	51008.9	<0.001	0.953
3	-20798.8	42945.6	**46216.7**	44076.1	**46880.7**	52857.9	**<0.001**	**0.981**
4	-20330.9	42459.7	46822.8	43967.7	47721.8	55681.1	<0.001	0.977
5	-19815.4	41878.8	47333.9	43764.2	48457.9	58409.1	<0.001	0.977
6	-19170	41037.9	47585	43300.7	48934.1	60877.1	<0.001	0.989

Application

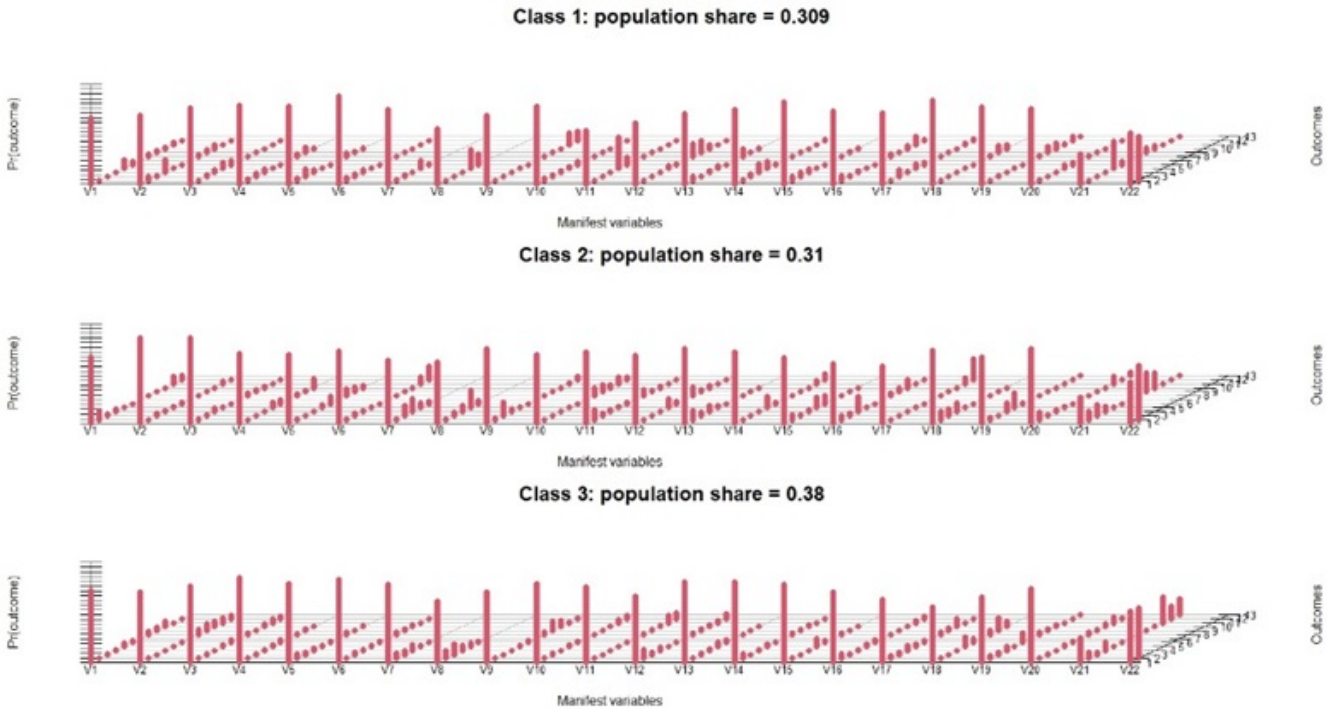


Figure 4: Latent Class Probabilities.

Application

Table 4: Machine Learning Results

Model(10 fold Cross Validation)	ACC (Train) 70%	ACC(Test)30%
LR	0.76	0.663
LDA	0.754	0.667
KNN	0.766	0.604
CART	0.872	0.811
NB	0.701	0.663
SVM	0.917	0.772
Xgboost	**1**	**0.951**
Adaboost	0.784	0.744
RF	0.947	0.854
NNET	0.916	0.765

Application

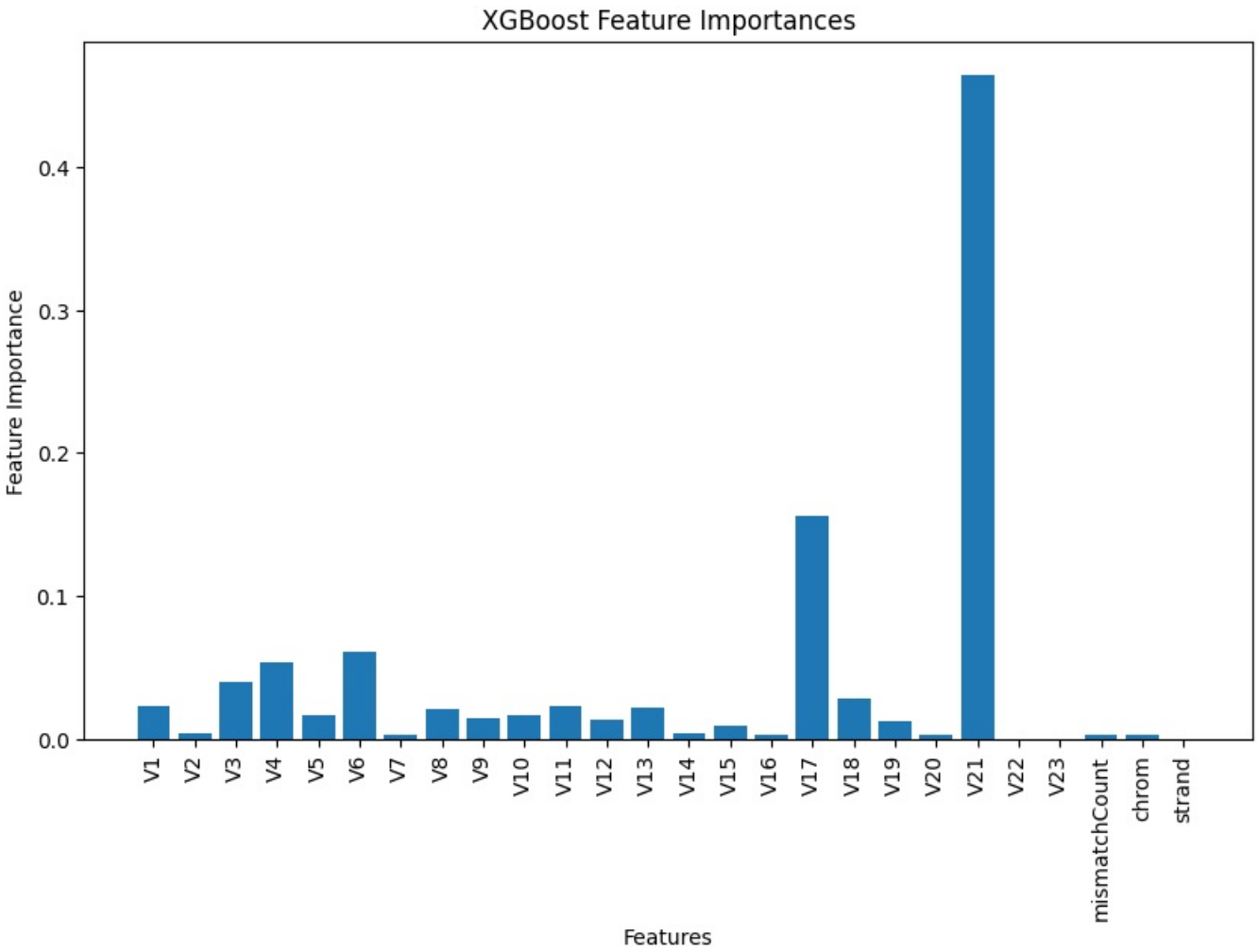


Figure 5: Feature Importance (Xgboost)

Discussion

Table 5: Literature Reviews

Study	Year	Model	Results
Vora et.al.	2022	Random Forest	%97 (ACC)
Chen et.al.	2018	Logistik Regression	%94 (ACC)
Aktaş et.al.	2019	Neural Networks	%96 (ACC)
Zhang et.al.	2022	Support Vector Machines	%98.2 (ACC)
Dhanjal et.al.	2020	Xgboost	%91.4 (ACC)
Trivedi et.al.	2020	Support Vector Machines	%98 (AUC)
Chuai et.al.	2018	Convolutional Neural Networks	%80.4 (AUC)
Abadi et.al.	2017	Random Forest	%96 (AUC)
Peng et.al.	2018	Random Forest	%99 (AUC)
Lazzarotto et.al.	2020	Random Forest	%99.5 (AUC)
Listgarten et.al.	2018	Boosted Regression Forests	%98 (AUC)
Doench et.al.	2016	Support Vector Machines and Logistic Regression	%80 (AUC)
Zhang et.al.	2019	Adaboost	%93.8 (AUC)

References

- Abadi, S., Yan, W. X., Amar, D., & Mayrose, I. (2017). A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS Computational Biology, 13(10), 1–24. https://doi.org/10.1371/journal.pcbi.1005807
- Acquah, H. D.-G. (2018). Weighted Average Information Criterion for Selection of an Asymmetric Price Relationship. Alanya Akademik Bakış, 2(2), 147–155. https://doi.org/10.29023/alanyaakademik.343737
- Aktaş, Ö., Doğan, E., & Ensari, T. (2019). Derin Öğrenmeyle CRISPR/CAS9 Hedefleme Tahmini. 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT 2019), 277 (1 Vol). Alpaydın, E. (2010). Introduction to Machine Learning (Second). The MIT Press.
- Asparouhov, T., & Muthén, B. (2014). Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using Mplus. Structural Equation Modeling, 21(3), 329–341. https://doi.org/10.1080/10705511.2014.915181

References

- Aurélien Géron. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. In O'Reilly Media. O'REILLY. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
- Barrangou, R., Fremaux, C., Deveau, H., Richardss, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). CRISPR provides against viruses in Prokaryotes. Science, 315(5819), 1709–1712. https://doi.org/10.1126/science.1138140
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. In Artificial Intelligence Review (Vol. 54, Issue 3). Springer Netherlands. https://doi.org/10.1007/s10462-020-09896-5
- Biemer, P. P. (2011). Latent Class Analysis of Survey Error (C. S. Mick P. Couper, Graham Kalton, J. N. K. Rao, Norbert Schwarz (ed.)). Wiley Series. https://doi.org/10.1002/9780470891155

References

- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. https://doi.org/10.1007/BF00746534
- Bolotin, A., Quinquis, B., Sorokin, A., & Dusko Ehrlich, S. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology, 151(8), 2551–2561. https://doi.org/10.1099/mic.0.28048-0
- Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345–370. https://doi.org/10.1007/BF02294361
- Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. Journal of Mathematical Psychology, 44(1), 62–91. https://doi.org/10.1006/jmps.1999.1277
- Bozok Çetintaş, V., Kotmakçi, M., & Tezcanli Kaymaz, B. (2017). From the Immune Response to the Genome Design; CRISPR-Cas9 System: Review. Turkiye Klinikleri Journal of Medical Sciences, 37(1), 27–42. https://doi.org/10.5336/medsci.2016-54153

Previous Publication

The CRISPR Journal
Volume 7, Number 3, 2024

DOI: 10.1089/crispr.2024.0002

The CRISPR Journal

RESEARCH ARTICLE

Unveiling Off-Target Mutations in CRISPR Guide RNAs: Implications for Gene Region Specificity

Ali Mertcan Kose,[1] Ozan Kocadagli,[2,*] Cihan Taştan,[3] Cagdas Aktan,[4] Onur Mert Ünaldı,[3] Elanur Güzenge,[3] and Hamza Emir Erdil[3]

Abstract

The revolutionary CRISPR-Cas9 technology has revolutionized genetic engineering, and it holds immense potential for therapeutic interventions. However, the presence of off-target mutations and mismatch capacity poses significant challenges to its safe and precise implementation. In this study, we explore the implications of off-target effects on critical gene regions, including exons, introns, and intergenic regions. Leveraging a benchmark dataset and using innovative data preprocessing techniques, we have put forth the advantages of categorical encoding over one-hot encoding in training machine learning classifiers. Crucially, we use latent class analysis (LCA) to uncover subclasses within the off-target range, revealing distinct patterns of gene region disruption. Our comprehensive approach not only highlights the critical role of model complexity in CRISPR applications but also offers a transformative off-target scoring procedure based on ML classifiers and LCA. By bridging the gap between traditional target-off scoring and comprehensive model analysis, our study advances the understanding of off-target effects and opens new avenues for precision genome editing in diverse biological contexts. This work represents a crucial step toward ensuring the safety and efficacy of CRISPR-based therapies, underscoring the importance of responsible genetic manipulation for future therapeutic applications.

Figure 6: Kose et. al, 2024, The CRISPR Journal

Previous Publication

Figure 7: Paper s Qr Code