**PhageAcr: Identification of anti-CRISPR proteins**

Christian Neitzel, Fernanda Vieira, Hugo Oliveira, Óscar Dias

Center of Biological Engineering, University of Minho, 4710-057 Braga, Portugal

**Introduction**

Bacteria and bacteriophages are engaged in a co-evolutionary arms race, driving the evolution of sophisticated defense mechanisms on both sides. Among these defenses, the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) system has emerged as a fascinating mechanism, enabling bacteria to fend off viral intrusion by precisely targeting and degrading viral DNA.

In response to the formidable CRISPR-Cas defense, bacteriophages have devised a potent countermeasure: Anti-CRISPR proteins (Acrs). These compact molecules, encoded by select bacteriophages, act as stealthy inhibitors, specifically targeting bacterial host immunity, particularly the CRISPR-Cas systems. By disabling CRISPR-Cas mechanisms, Acrs ensure the successful invasion and replication of bacteriophages within their bacterial hosts.

Understanding the intricacies of Anti-CRISPR systems is crucial for advancing genetic engineering and biotechnology. This article aims to contribute to this understanding by fulfilling two main objectives. Firstly, it seeks to build a dataset of Acr protein sequences utilizing available datasets and NCBI queries. Secondly, it aims to initiate the exploration of machine learning approaches to enhance the accuracy of Acr predictions.

By dissecting the mechanisms through which CRISPR-Cas defends bacteria and elucidating the role of Acrs as mediators, invaluable insights can be gleaned to propel innovation and breakthroughs in these fields. The development of bioinformatic tools, leveraging Machine Learning models, refining Acr nomenclature, and tapping into existing databases of known Anti-CRISPR proteins are pivotal steps in this pursuit. These advancements promise to not only augment our understanding of Acrs, but also lay the groundwork for further exploration and application in microbial ecology and genome engineering.

1. **Bacterial defense system**

Bacteriophages (phages) are the most abundant viruses on the planet, capable of infecting and replicating within bacteria by attaching to specific receptors on the bacterial surface and injecting their genetic material into the host cell. Following infection, phages may undergo either a lytic cycle, during which they replicate within the host cell and ultimately cause the cell to lyse, releasing new phage particles, or a lysogenic cycle, in which the phage genome integrates into the bacterial chromosome. In the lysogenic cycle, the integrated phage genome can replicate along with the bacterial chromosome and remain dormant until it is induced to enter the lytic cycle and produce new phage particles.

Phages play crucial roles in regulating bacterial populations, in nutrient cycling and in maintaining microbial diversity in various ecosystems, ranging from oceans, soil and within other living organisms. Additionally, phages have garnered attention for their potential applications in biotechnology, including phage therapy for combating bacterial infections, and as tools for genetic engineering that has been demonstrating promising developments in the coming years (Haq et al., 2012).

Being phages the most predominant and diverse lifeform in the planet, , bacteria are evolutionarily pressured to develop a diversity of defenses in order to survive. Previous works allude to this line of defense as constituting a prokaryotic immune system (Bernheim & Sorek, 2020). Of these countermeasures, the most known are: the restriction modification (RM) system, which degrades viral DNA by cleaving it into fragments so that it becomes susceptible to degradation by endonucleases; the abortive infection (Abi) system, which orchestrates the programmed death of the infected host to prevent the phage replication cycle and finally, the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR-Cas) system.

The CRISPR defense system has garnered significant attention due to its genome editing properties which are of great importance in the area of genetic engineering.

CRISPR-Cas or CRISPR are a family of adaptive immune systems that function through the recognition and degradation of viral nucleic acids. This system is composed of two sections: a CRISPR array and a cluster of genes that encode Cas proteins.

The CRISPR array contains direct repeat sequences interspersed with variable-length “spacers”. These spacers represent fragments of viral DNA previously encountered by the bacteria, having been acquired when the viral DNA sequence they originated from was cleaved and modified by a Cas protein. The spacers will then allow the bacteria to “remember” these nucleic acids, optimizing the defense mechanisms should similar phages attempt to infect the bacteria.

CRISPR-Cas systems are characterized by three primary stages: adaptation, expression and interference.

The adaptation stage occurs when Cas proteins recognize a specific motif, known as a Protospacer-Adjacent Motif (PAM), on target DNA and remove a segment referred to as the protospacer. This process enables the integration of the protospacer DNA into the CRISPR array, serving as a spacer for future immunity. Some CRISPR-Cas systems employ an alternative mechanism, acquiring spacers from RNA through reverse transcription, thereby enhancing their adaptive capabilities.

In the Expression stage, the CRISPR array is transcribed into a single pre-CRISPR RNA (pre-crRNA), which undergoes processing to yield mature CRISPR RNAs (crRNAs). These crRNAs contain spacer sequences along with flanking repeats, which are essential for target recognition and interference.

During the Interference stage, crRNA guides the CRISPR-Cas complex to recognize and cleave the protospacer or similar sequences within the genome of invading entities. This precise targeting results in the cleavage and inactivation of the invader's genetic material by Cas nucleases, providing a robust defense against viral infections and gene transfer events.

1. **CRISPR types and subtypes**

Like other defense mechanisms, CRISPR-Cas systems possess a wide variety of Cas protein sequences, gene compositions and genomic structures. This diversity suggests that these systems evolved and diversified across different species and environments, possibly as responses to the various pathogens encountered by the bacteria and archaea in which they existed. To further understand this diversity, researchers developed a classification of CRISPR-Cas systems.

CRISPR-Cas systems are classified based on their evolutionary relationships through comparisons of gene compositions, locus architectures, sequence similarity-based clustering and phylogenetic analysis of conserved Cas proteins (Makarova et al., 2015), the results showed that two classes of CRISPR-Cas systems can be defined based on significant differences in the architectures of the effector modules involved in crRNA processing and interference.

Class 1 systems feature effector modules comprising multiple Cas proteins that collaborate to form a complex binding with crRNA, working in tandem to bind and process the target. In contrast, class 2 systems are characterized by a single, multidomain crRNA-binding protein, which serves a functionally similar role to the entire effector complex found in class 1 systems (Makarova et al., 2020).

Within these two classes, we can further classify distinct types based on the presence of unique signature Cas genes (Makarova & Koonin, 2015), (S. Makarova et al., 2011), such as Cas3 being unique in type I, Cas9 in type II, Cas10 in type III, etc.

Class 1 is composed of 3 types (types I, III and IV) and 17 subtypes whereas class 2 is composed of 3 types (types II, V, VI) and 33 subtypes (S. Makarova et al., 2011).

* 1. **Tools to detect CRISPR systems**

To help identify CRISPRs, several tools were developed to assist in the identification and classification of CRISPR-Cas systems, some of which being CRISPRloci, CRISPRFinder and CRISPRCasFinder.

CRISPRloci utilizes Machine Learning algorithms to predict and assess CRISPR loci accurately, offering precise annotation of Cas genes and providing comprehensive information on CRISPR array characteristics, including orientation, leader sequence, and Cas subtype classification, streamlining the process of CRISPR-Cas system characterization (Alkhnbashi et al., 2021).

CRISPRFinder is a web tool that allows the identification of CRISPR-like structures within genomic sequences, characterized by its ability to define Direct Repeats (DRs) and extract spacers, as well as its inclusion of smaller CRISPR candidates.

While previous tools often dismissed smaller CRISPR suspects (one or two motifs in size), CRISPRFinder’s inclusive approach sheds light on CRISPR evolution and spread. Despite the possibility of false candidates being wrongly included, CRISPRFinder addresses this challenge by labelling them as “questionable” paving the way for future refinement and accurate identification (Grissa et al., 2007).

CRISPRCasFinder, an advanced iteration of CRISPRFinder, integrates the identification of both CRISPR arrays and Cas proteins. With an improved CRISPR array detection tool, it demonstrates superior performance in identifying Cas proteins, aligning seamlessly with the latest classification schemes (Couvin et al., 2018).

Notably, this tool not only enhances performance but also consolidates functions from other tools for comprehensive CRISPR identification.

1. **Phage Anti-CRISPR**

Bacteria and archaea evolve anti-viral defenses, triggering evolutionary responses from phages to evade host immune systems, sparking an ongoing arms race. Phages deploy Anti-CRISPR proteins (Acrs) as natural inhibitors of CRISPR-Cas systems, halting bacterial immune responses. Injected alongside viral DNA during infection, Acrs target and block Cas proteins involved in DNA cleavage, disabling CRISPR-Cas mechanisms and enabling unhindered infection (Bondy-Denomy et al., 2013).

CRISPR-Cas has been harnessed for gene editing and synthetic gene circuit construction. Since Acrs act as OFF-switches for CRISPR-Cas activity, they prove to be beneficial in the development of CRISPR-Cas based biotechnological tools, holding a promise for enhancing gene editing techniques and addressing potential challenges in eukaryotic synthetic gene circuits (Yu & Marchisio, 2020).

* 1. **Databases for AntiCRISPR**

To accelerate and broaden the exploration of Acrs, user-friendly and regularly updated databases have been established with Anti-CRISPRdb being a prominent example. Leveraging Python scripts, this database extracts data from PubMed and Google Scholar, followed by sequence alignment through BLAST from NCBI. Its primary goal is to streamline researchers' efforts by providing comprehensive functions for searching, browsing, downloading and uploading data on anti-CRISPR proteins. With accessible information encompassing family, function, source species, verification status, comments and references, researchers can efficiently navigate through the database. The intuitive interface, coupled with robust search capabilities and BLAST integration, facilitates swift access to Acr proteins and their potential candidates, thus fostering accelerated research progress in the field (Dong et al., 2018).

* 1. **Tools to detect AntiCRISPR**

Identifying Acrs presents a noteworthy challenge, largely due to the variability in their amino acid sequences, reflecting the diverse nature of CRISPR-Cas systems found in bacteria and archaea. Their compact size and extensive evolutionary diversity add further complexity to detection, hindering even the most sophisticated sequence analysis methods. Traditionally, bioinformatic tools for Acr identification relied on guilt-by-association and self-targeting methods. However, Machine Learning approaches for Acr prediction models can accelerate and enhance the identification of candidate Acr families. This can be effectively achieved by providing the model with a database of confirmed Acrs, to learn their distinguishing characteristics, and a database of non-Acrs to discern what to exclude, thereby equipping the model with a clear understanding of the features that differentiate Acrs from non-Acrs (Gussow et al., 2020).

AcrFinder (available at <http://bcb.unl.edu/AcrFinder>), integrates a multifaceted and well-accepted approach, combining homology search, guilt-by-association (GBA) and CRISPR-Cas self-targeting spacers, allowing for increased likelihood of identifying Acrs (Yi et al., 2020).

Homology search identifies Acr homologs based on sequence similarity and GBA identifies Aca proteins co-localized with Acrs, indicating their functional association. Meanwhile the analysis of CRISPR-Cas self-targeting spacers provides additional validation for the presence of Acr-Aca operons, indicating the interactions between Acr proteins and the CRISPR-Cas systems.

With Anti-CRISPRDB focusing on cataloging experimentally characterized Acr proteins and their homologs, providing quick access to curated data (Dong et al., 2018), and CRISPRminer offering advanced functionalities for analyzing CRISPR-Cas systems, predicting new Acr candidates, and providing extensive analysis tools and prediction capabilities (Zhang et al., 2018). Both resources serve as valuable tools for detecting anti-CRISPR proteins. Each with unique advantages, and they can complement each other in the study of anti-CRISPR proteins and CRISPR-Cas systems.

With the continuous discovery of new Acr protein families, establishing a standardized nomenclature has become imperative for ensuring clarity and consistency among researchers. Several efforts have been undertaken to develop such a nomenclature, aiming to streamline the naming process and facilitate communication within the scientific community. Notable contributions include the works cited in references (Pawluk et al., 2016), (Bondy-Denomy et al., 2015), which laid the groundwork for a unified approach to naming Acrs.

Building upon this foundation, some researchers have further advanced the initiative by creating a Google Doc database adhering to the standardized nomenclature (Bondy-Denomy et al., 2018). This database, accessible via (<https://tinyurl.com/anti-CRISPR>), serves as a comprehensive repository for cataloging newly identified Acr proteins and ensuring conformity to established naming conventions (Bondy-Denomy et al., 2018).

XX

Objectives

* Build a dataset
* Benchmark of developed tools

Methods

**Bibliography**

Alkhnbashi, O. S., Mitrofanov, A., Bonidia, R., Raden, M., Tran, V. D., Eggenhofer, F., Shah, S. A., Öztürk, E., Padilha, V. A., Sanches, D. S., de Carvalho, A. C. P. L. F., & Backofen, R. (2021). CRISPRloci: Comprehensive and accurate annotation of CRISPR–Cas systems. *Nucleic Acids Research*, *49*(W1), W125–W130. https://doi.org/10.1093/nar/gkab456

Bernheim, A., & Sorek, R. (2020). The pan-immune system of bacteria: Antiviral defence as a community resource. *Nature Reviews Microbiology*, *18*(2), 113–119. https://doi.org/10.1038/s41579-019-0278-2

Bondy-Denomy, J., Davidson, A. R., Doudna, J. A., Fineran, P. C., Maxwell, K. L., Moineau, S., Peng, X., Sontheimer, E. J., & Wiedenheft, B. (2018). A Unified Resource for Tracking Anti-CRISPR Names. *The CRISPR Journal*, *1*(5), 304–305. https://doi.org/10.1089/crispr.2018.0043

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M. F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K. L., & Davidson, A. R. (2015). Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature*, *526*(7571), 136–139. https://doi.org/10.1038/nature15254

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L., & Davidson, A. R. (2013). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, *493*(7432), 429–432. https://doi.org/10.1038/nature11723

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, *46*(Web Server issue), W246–W251. https://doi.org/10.1093/nar/gky425

Dong, C., Hao, G.-F., Hua, H.-L., Liu, S., Labena, A. A., Chai, G., Huang, J., Rao, N., & Guo, F.-B. (2018). Anti-CRISPRdb: A comprehensive online resource for anti-CRISPR proteins. *Nucleic Acids Research*, *46*(Database issue), D393–D398. https://doi.org/10.1093/nar/gkx835

Grissa, I., Vergnaud, G., & Pourcel, C. (2007). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(Web Server issue), W52–W57. https://doi.org/10.1093/nar/gkm360

Gussow, A. B., Park, A. E., Borges, A. L., Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Bondy-Denomy, J., & Koonin, E. V. (2020). Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nature Communications*, *11*(1), 3784. https://doi.org/10.1038/s41467-020-17652-0

Haq, I. U., Chaudhry, W. N., Akhtar, M. N., Andleeb, S., & Qadri, I. (2012). Bacteriophages and their implications on future biotechnology: A review. *Virology Journal*, *9*, 9. https://doi.org/10.1186/1743-422X-9-9

Makarova, K. S., & Koonin, E. V. (2015). Annotation and Classification of CRISPR-Cas Systems. *Methods in Molecular Biology (Clifton, N.J.)*, *1311*, 47–75. https://doi.org/10.1007/978-1-4939-2687-9\_4

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., … Koonin, E. V. (2015). An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology*, *13*(11), 722–736. https://doi.org/10.1038/nrmicro3569

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Scott, D., Shah, S. A., Siksnys, V., Terns, M. P., Venclovas, Č., White, M. F., Yakunin, A. F., … Koonin, E. V. (2020). Evolutionary classification of CRISPR-Cas systems: A burst of class 2 and derived variants. *Nature Reviews. Microbiology*, *18*(2), 67–83. https://doi.org/10.1038/s41579-019-0299-x

Pawluk, A., Amrani, N., Zhang, Y., Garcia, B., Hidalgo-Reyes, Y., Lee, J., Edraki, A., Shah, M., Sontheimer, E. J., Maxwell, K. L., & Davidson, A. R. (2016). Naturally Occurring Off-Switches for CRISPR-Cas9. *Cell*, *167*(7), 1829-1838.e9. https://doi.org/10.1016/j.cell.2016.11.017

S. Makarova, K., H. Haft, D., Barrangou, R., J. J. Brouns, S., Charpentier, E., Horvath, P., Moineau, S., J. M. Mojica, F., I. Wolf, Y., Yakunin, A. F., van der Oost, J., & V. Koonin, E. (2011). Evolution and classification of the CRISPR-Cas systems. *Nature Reviews. Microbiology*, *9*(6), 467–477. https://doi.org/10.1038/nrmicro2577

Yi, H., Huang, L., Yang, B., Gomez, J., Zhang, H., & Yin, Y. (2020). AcrFinder: Genome mining anti-CRISPR operons in prokaryotes and their viruses. *Nucleic Acids Research*, *48*(W1), W358–W365. https://doi.org/10.1093/nar/gkaa351

Yu, L., & Marchisio, M. A. (2020). Types I and V Anti-CRISPR Proteins: From Phage Defense to Eukaryotic Synthetic Gene Circuits. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.575393

Zhang, F., Zhao, S., Ren, C., Zhu, Y., Zhou, H., Lai, Y., Zhou, F., Jia, Y., Zheng, K., & Huang, Z. (2018). CRISPRminer is a knowledge base for exploring CRISPR-Cas systems in microbe and phage interactions. *Communications Biology*, *1*, 180. https://doi.org/10.1038/s42003-018-0184-6