CRISPR-Act: AI-guided Prediction of a CRISPR Kill Switch Under Diverse Physiological Conditions

Rebecca Weinberg[1,4*] (rweinberg@anl.gov), Carla M. Mann,[2,4] Gyorgy Babnigg,[1] Sara Forrester,[1] Stephanie Greenwald, [1] Peter E. Larsen,[1] Sarah Owens, [1] Marie-Francoise Gros,[1,3] Philippe Noirot,[1,3] Arvind Ramanathan,[2] and **Dionysios A. Antonopoulos**[1]

[1]Biosciences Division, Argonne National Laboratory, Lemont, IL; [2]Data Science and Learning Division, Argonne National Laboratory, Lemont, IL; [3]National Research Institute for Agriculture, Food and Environment (INRAE), France; [4]These authors contributed equally

Project Goals: The long-term goal for this Project is to realize secure biodesign strategies for microbial systems that operate in the dynamic abiotic and biotic conditions of natural environments, thus enabling systems-level and rational biological design for field use. There are several key challenges to incorporating safeguard systems at the design stage including: (1) lack of knowledge for how well safeguards operate across the broad set of environmental and physiological conditions that an organism experiences; (2) a need to integrate the safeguard with other cellular components so that it can sense and recognize specific signals from the intracellular or extracellular environment, and mediate a response; and (3) a need for rapid and reliable methods to engineer and optimize the biological components for safeguard construction and functional integration. To address these challenges, we propose to utilize recent advances in the fields of synthetic biology, artificial intelligence (AI), and automation, which are creating the conditions for a paradigm shift in our understanding of the ways that cellular function can be designed at the level of bacterial communities.

The development of genetically modified and engineered organisms necessitates the creation of secure and efficient biocontainment systems to prevent these organisms from escaping the laboratory and endangering the environment, public health, and public perception of scientific research. Safeguards based on controlled activation of a self-targeting CRISPR/Cas9 "self-destruct" mechanism that activates outside of laboratory conditions are transferable between different organisms, cost-effective, and relatively easy to implement in comparison to other safeguard systems. However, the variability of CRISPR/Cas cleavage efficiency within a genome (and even within the same gene) [1, 2] represents a challenge in choosing efficient self-targeting guide RNAs (gRNAs) for self-destruction, particularly as that variability is heightened under different environmental conditions. We have developed a machine-learning prediction method, CRISPRAct, that models this behavior across different environmental conditions to assist in identifying gRNAs for use in the proposed secure biosystem.

We hypothesized that dynamic gene expression responses to varying physiological conditions would influence the cell killing activity of the CRISPR/Cas9 system. We created a library of gRNAs composed of ~180,000 gRNAs corresponding to sites throughout the *Escherichia coli* MG1655 genome and control gRNAs (~20,000) that did not match. Using this library, we conducted screens for killing activity in three physiological conditions: rich (LB) media in exponential growth; defined (M9) media in exponential growth; and rich media in stationary phase. We then captured sequence data representing the transcriptomes corresponding to these conditions, and the killing activity of the gRNAs over time. Using our libraries, we identified ~6,000 guides that were statistically overrepresented (via ANOVA testing) for

physiology-specific function. A small subset of gRNAs (174) were "outlier switches" that exhibited outstanding killing activity in a specific physiological condition, while their high prevalence in another condition was consistent with providing a growth advantage. The library screens generated an extensive dataset of ~530,000 data points used to develop CRISPRAct to interrogate features influencing gRNA killing activity.

The CRISPRAct model combines a state-of-the-art natural language processing (NLP) model based on the Google AI ALBERT architecture [3] with conventional neural network (NN) models to predict the efficiency of gRNAs under various conditions. The NLP model treats genomic context as a machine-interpretable "language" by training embeddings on a representative set of twelve *E. coli* genomes [4] divided into "sentences" of seven "words" that each consist of three bases, and then fine-tuning on the gRNA activity dataset. The NN models use a set of 428 features including gRNA positional and physicochemical properties [1], as well as the *E. coli* growth phase in the form of optical density, and environmental factors including media growth concentrations. The results of these models are combined through a polynomial regression to predict the percent change in *E. coli* cell population after Cas9 induction. CRISPRAct achieves a Mean Square Error of 0.13 and Spearman Correlation Coefficient of 47.31% on a test dataset of 71,228 gRNAs. This level of model performance is comparable to other efforts to predict gRNA activity. However, due to our approach of leveraging physiological conditions to train models, CRISPRAct offers a more resilient basis for transfer learning in novel organisms and environmental conditions without having to undergo a costly re-training process. This strategy will decrease the time and effort needed by other researchers to identify gRNAs with desired behaviors in various physiological conditions.

References

1. Guo, J., et al., *Improved sgRNA design in bacteria via genome-wide activity profiling*. Nucleic Acids Res, 2018. 46(14): p. 7052-7069.
2. Gutierrez, B., et al., *Genome-wide CRISPR-Cas9 screen in E. coli identifies design rules for efficient targeting*. bioRxiv, 2018: p. 308148.
3. Lan, Z., et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. arXiv, 2019: 1909.11942
4. Abram, K., et al., *What can we learn from over 100,000 Escherichia coli genomes?* bioRxiv, 2020

This Project is funded by the Biological Systems Science Division's Genomic Science Program, within the U.S Department of Energy, Office of Science, Biological and Environmental Research.