CRISPR-Act: AI-guided Prediction of a CRISPR Kill-switch Across Physiological Contexts

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 together pave the way 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 engineered organisms necessitates the creation of secure and efficient biocontainment systems to safely contain such organisms and thus protect the environment, public health, and public perception of scientific research. Microbial safeguards based on controlled activation of a self-targeting CRISPR/Cas9 “self-destruct” mechanism are transferable between different organisms, cost-effective, and relatively easy to implement. 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 gRNA efficiency also varies under different environmental conditions. We have developed a machine-learning prediction method, CRISPRAct, that predicts gRNA efficiency across different environmental conditions to assist in identifying candidate gRNAs for use in secure biosystems.

We hypothesized that dynamic gene expression responses to varying physiological conditions would influence the cell-killing activity of the CRISPR/Cas9 system. To test this hypothesis, we screened the activity of a library of 180,000 gRNAs spanning the *E. coli* MG1655 genome, and compared cell-killing activity to control sequences with no genomic matches (~20,000). We used the log2 fold change in cell population between time points as a proxy for gRNA cutting activity. To assess the influence of physiological conditions on gRNA cutting, screens were conducted as time courses in three growth conditions: rich media in exponential growth (LB-E); defined media in exponential growth (M9-E); and rich media in stationary phase (LB-S).

In our initial library screens, we identified ~6,000 guides that were statistically overrepresented (via ANOVA testing) for physiology-specific functions. A small subset of

gRNAs (174) were "outlier switches" that were statistical outliers in their outstanding killing activity in one or more physiological conditions, but were also statistical outliers in their lack of activity under the other physiological condition(s). Additionally, there is a marked difference in GC content between guides that efficiently kill in the M9 media versus the LB media: guides that have low activity in rich media (LB) but high activity in minimal media (M9) are GC rich, while guides that have high activity in rich media low activity in minimal media are very AT rich. Further, these "switch" guides tend to localize to specific areas in the genome – there are six regions that active rich media/inactive minimal media guides localize to, while inactive rich/active minimal media guides localize to three narrow regions and two broad swathes within the genome. We also identified a small but statistically significant correlation between the number of sites a gRNA can target in the genome, and the gRNA's cutting efficiency, and correlations of varying strengths between the proximity of a gRNA target site to nucleoid-associated protein binding motifs in the primary genomic sequence. Collectively, the initial round of 200k library screens generated an extensive dataset of more than 530,000 data points used to develop CRISPRAct.

The CRISPRAct model is a two-part model that combines a natural language processing (NLP) model with a neural network (NN) model. The NLP model treats genomic sequences as a machine-interpretable "language", while the NN model features now incorporate gRNA positional and physicochemical properties (including nucleoid-associated protein binding motif proximity) with environmental conditions. The outputs of these models are combined through polynomial regression to predict the percent fold change of guide prevalence after Cas9 induction, as a proxy for the gRNA activity. CRISPRAct achieves a Mean Absolute Error of 0.39 and Spearman Correlation Coefficient of 66.4%, beating the correlation from the previous state of the art (in a single environmental condition) by about 12%. Importantly, CRISPRAct exhibited a comparable Mean Absolute Error across physiological conditions- 0.51 in LB-E, 0.52 in LB-S and 0.54 in M9-E. CRISPRAct is thus, to our knowledge, the first gRNA activity predictor capable of predicting behavior under different environmental conditions. We are currently assessing reproducibility of our screens. Empirically assessed correlations between screens varied by physiological condition, which informs our strategy to develop transfer learning. Ultimately, we anticipate that leveraging these physiological variations while training CRISPRAct will improve the robustness of our models and reduce costly retraining time as we move into novel genomic contexts.

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.

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.