Machine Learning Guided Design of Safeguards That Operate Under Various Bacterial Physiologies

C.M. Mann[1], R. Weinberg[2], S. Forrester[2], G. Babnigg[2], PE. Larsen[2], MF. Gros[2], **A. Ramanathan**[1], and **P. Noirot**[2]* (pnoirot@anl.gov)

[1] Data Science and Learning Division, Argonne National Laboratory, Lemont, Il; [2] Biosciences Division, Argonne National Laboratory, Lemont, Il.

Project Goals: Our overarching goal is to build and train machine learning models that can predict the activity of the complexes between the CRISPR Cas9 nuclease and guide RNAs (Cas9-gRNA) in various genomes. To reach this goal, we are developing a machine learning approach that predicts the activity of Cas9-gRNA complexes based on features from the targeted DNA sequences and from the context of these targets, such as genome organization, annotation, and gene expression. The rules learned by the model will be transferred to another genomic context and validated experimentally. Finally, gRNAs that are highly efficient at killing bacteria will be used to develop safeguard systems.

The growing deployment of engineered organisms for environmental, bioenergy, and industrial applications represents an ever-increasing risk of releasing these organisms in the environment. To limit this risk, efficient biocontainment systems must be developed to prevent the survival of even a small number of released genetically engineered organisms. A safeguard system based on the controlled activation of a CRISPR nuclease, which breaks the chromosome at multiple targeted sites and kills the cell, is a design that is portable between organisms, including in non-model bacterial species of relevance in many environmental and biotechnological processes of interest to DOE. However, Cas9-gRNA complexes exhibit cleavage efficiencies that vary considerably along the genome [1, 2], limiting the use of CRISPR/Cas9 in safeguard systems.

We hypothesize that physiological conditions drive key factors that influence Cas9-gRNA activity in bacteria. To test this hypothesis, we are developing a machine learning approach, called CRISPRAct, to predict the activity of Cas9-gRNA complexes based on features from the targeted DNA sequences and from the context of these targets, such as genome organization, annotation, and gene expression. We are generating genome-wide Cas9-gRNA activity profiles in *E. coli* by screening a library of ~ 200,000 gRNAs under different physiological conditions. The datasets will be used to train and validate our CRISPRAct model to predict condition-specific Cas9-gRNA activity along the *E. coli* chromosome. The gRNAs predicted to be highly efficient at killing *E. coli* in different physiological conditions will be used to build safeguards.

The prediction of Cas9-gRNA activity not only as a function of DNA sequence but also including specific features from genome context and environmental conditions goes beyond how current models predict optimal gRNA design. These additional features will be key to apply a transfer learning approach to predict gRNA-Cas9 activity in other bacterial genomes. In this project, we will apply transfer learning to predict Cas9-gRNA activity in *Pseudomonas fluorescens*. Our models will be made publicly available through an integrated Python software package. Our results and tools will open the way to the portable design of CRISPR-based safeguards in multiple non-model bacterial species.

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 <em>E. coli</em> identifies design rules for efficient targeting.* bioRxiv, 2018: p. 308148.

This work is funded by the Department of Energy, office of Biological and Environmental Research.