Manuscript Title

This manuscript was automatically generated on February 8, 2023.

Authors

- **John Doe**
  iD XXXX-XXXX-XXXX-XXXX · johndoe · johndoe · @johndoe@mastodon.social
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  iD XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

How can the efficiency and specificity of CRISPR-based tools be maximized? To address this problem, this paper proposes a machine learning-based approach for gRNA design for multiple CRISPR-Cas systems. We develop a Markdown-based publishing platform, Manubot, to facilitate the efficient administration of CRISPR-based tools. This platform provides a comprehensive set of design rules for various CRISPR-Cas systems, enabling the prediction of gRNAs with high on-target efficiency and off-target specificity.

This work investigates the design rules for CRISPR interference (CRISPRi) to understand its potential for functional interrogation, pathway manipulation, and genome-wide screens in bacteria. To address this research problem, a state-of-art predictive machine learning model is developed to predict guide silencing efficiency in bacteria. This model leverages the advantages of feature engineering, data integration, interpretable AI, and automated machine learning. Results show that gene-specific effects, such as gene expression level, make a dominant contribution to the extent of depletion in multiple CRISPRi genome-wide essentiality screens in Escherichia coli. This observation allows for the confounding gene-specific effects to be segregated using the mixed-effect random forest (MERF) model to provide a better estimate of guide efficiency. The MERF model outperforms existing tools in the independent datasets, including small-scale fluorescence-based and miller assays. Design rules for robust gene silencing are extracted, such as the preference for cytosine and disfavoring for guanine and thymine within and around the PAM sequence. This work is concluded with a web-based tool, freely accessible at www.ciao.helmholtz-hiri.de, that incorporates the MERF model.

This paper presents a machine learning approach to effectively develop prediction algorithms for CRISPR-Cas systems. The workflow includes three principle steps: accommodating the feature set for the CRISPR-Cas system or technique; optimizing a machine learning model using automated machine learning; and explaining the model using interpretable AI. This workflow was applied to analyze three different CRISPR-Cas genome-wide screens, and the results revealed various design rules for efficient editing, target expression level as a main determinant of activation of Cas13-induced immunity, and robust antimicrobial activity across K. pneumoniae strains. The proposed approach provides a blueprint for the development of prediction algorithms that are robust across organisms and CRISPR-Cas techniques.

This thesis presents a machine learning approach to accurately predict the CRISPRi guide efficiency in bacteria. This systematic exploration of the determinants of efficient silencing and guide designs has led to a robust model which can be applied to other bacteria and CRISPR-Cas systems. Furthermore, the analysis of independent CRISPR-Cas screens using this model provides insights into the design rules and mechanisms of the CRISPR-Cas systems. The results of this research demonstrate that machine learning can be successfully applied to gain a deeper understanding and broader application of CRISPR-Cas systems.

References