Improved prediction of bacterial CRISPRi guide efficiency through data integration and automated machine learning

This manuscript was automatically generated on February 8, 2023.

Authors

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

- **Jane Roe** ✉
  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

CRISPRi is an effective tool for gene knockdown. However, the efficiency of CRISPRi varies greatly at different targets. The current methods for predicting CRISPRi efficiency are not accurate enough. This paper proposes a new method for predicting CRISPRi efficiency based on data integration and automated machine learning. The new method combines three models and uses an automated machine learning framework to generate a hybrid model. The new method is shown to outperform the current methods.

The research problem: CRISPR interference (CRISPRi), an alternative gene silencing technique using a catalytically dead Cas protein to interfere with transcription, is one of these CRISPR-Cas systems. The under-investigated design rules for CRISPRi limit its promising application for functional interrogation, pathway manipulation, and genome-wide screens in bacteria.

The solution proposed: In this work, we first developed a state-of-art predictive machine learning model for guide silencing efficiency in bacteria leveraging the advantages of feature engineering, data integration, interpretable AI, and automated machine learning. We systematically investigated the influential factors that attribute to the extent of depletion in multiple CRISPRi genome-wide essentiality screens in Escherichia coli and demonstrated the surprising dominant contribution of gene-specific effects, such as gene expression level. This observation allowed us to segregate the confounding gene-specific effects using the mixed-effect random forest (MERF) model to provide a better estimate of guide efficiency, together with the improvement led by integrating multiple screens. The MERF model outperformed existing tools in the independent datasets, including small-scale fluorescence-based and miller assays involving only dozens of gRNAs and a high-throughput saturating screen involving 750 gRNAs. We next interpreted the predictive model to extract the design rules for robust gene silencing, such as the preference for cytosine and disfavoring for guanine and thymine within and around the PAM sequence. We further incorporated the MERF model in a web-based tool that is freely accessible at www.ciao.helmholtz-hiri.de.

The performance of the gRNA design tool optimized for CRISPRi in eukaryotes was far from satisfying, questioning the robustness of prediction algorithms across organisms. In addition, the CRISPR-Cas systems exhibit diverse mechanisms albeit sharing similarities. The captured predictive patterns from one dataset are at risk of poor generalization when applied across organisms and CRISPR-Cas techniques. To fill the gap, the machine learning approach that I present for CRISPRi could serve as a blueprint for the effective development of prediction algorithms. The explicit workflow includes three principle steps: 1) accomodating the feature set for the CRISPR-Cas system or technique; 2) optimizing a machine learning model using automated machine learning; 3) explaining the model using interpretable AI. I applied this workflow to analyze three other CRISPR-Cas genome-wide screens. From the CRISPR base editor essentiality screen in E. coli, I determined the PAM preference and sequence context in the editing window for efficient editing, such as A at the 2nd position of PAM, A/TT/TG downstream of PAM, and TC at the 4th to 5th position of gRNAs. From the CRISPR-Cas13a screen in E. coli, in addition to the strong correlation with the guide depletion, the target expression level was the strongest predictor in the model, supporting it as a main determinant of the activation of Cas13-induced immunity and better characterizing the CRISPR-Cas13 system. From the CRISPR-Cas12a screen in Klebsiella pneumoniae, I extracted the design rules for robust antimicrobial activity across K. pneumoniae strains and provided a predictive algorithm for gRNA design, facilitating CRISPR-Cas12a as an alternative technique to tackle antibiotic resistance.

We propose a machine learning approach to predict the CRISPRi guide efficiency in bacteria, providing insights into the determinants of efficient silencing and guide designs. The systematic exploration has

led to a robust machine learning approach for effective model development in other bacteria and CRISPR-Cas systems. Applying the approach in the analysis of independent CRISPR-Cas screens not only sheds light on the design rules but also the mechanisms of the CRISPR-Cas systems. Together, we demonstrate that applied machine learning paves the way to a deeper understanding and a broader application of CRISPR-Cas systems.

References