DOI 10.18699/SBB-2022-670

Computational identification of Anti-CRISPR proteins by using information extracted from the primary structures

Sidra L.[1], Saiqa A.[2], Maryum B.[1], Wajid A.A.[1]*

[1] *Computational Biology and Data Analysis Lab., Department of Computer Sciences & Information Technology, King Abdullah Campus, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan*

[2] *Biotechnology Lab., Department of Zoology, King Abdullah Campus, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan*

** wajidarshad@ajku.edu.pk; wajidarshad@gmail.com*

Key words: Anti-CRIPR, proteins, machine learning, sequence information, predictive modeling

Motivation and Aim: The CRISPR-CAS technique has revolutionized genome editing, resulting in intrinsically engineered disease-resistant crops, livestock & other products. It has immense application in the personalized treatment of human diseases such as cancer, cystic fibrosis, and muscular dystrophy, and can be effective against infectious diseases such as malaria and HIV. The problem with the CRISPR-CAS system is that it lacks an off-switch to turn off Cas9's activity, which leads to off-target edits and mutations, resulting in adversity [1]. Anti-CRISPR (ACR) proteins proved to be natural off-switching-inhibitors for CAS's activity. Due to the immutable significance of ACR's a more consistent, quicker, and accurate approach is needed for discovering the Anti-CRISPR proteins to increase the rate and volume of discovered ACR's.

Methods and Algorithms: We have employed various machine leaning algorithms such as large margin (Support Vector Machines), bagging based ensembles (Random Forest), and boosting based ensembles (Gradient boosting Machine). We used sequence derived structural descriptors of protein sequences to train and test different models. Various biological significant evaluation coupled with machine learning based matrices have been employed to gauge the generalization performance of our proposed solution. A generic framework of this study is shown in Fig. 1.

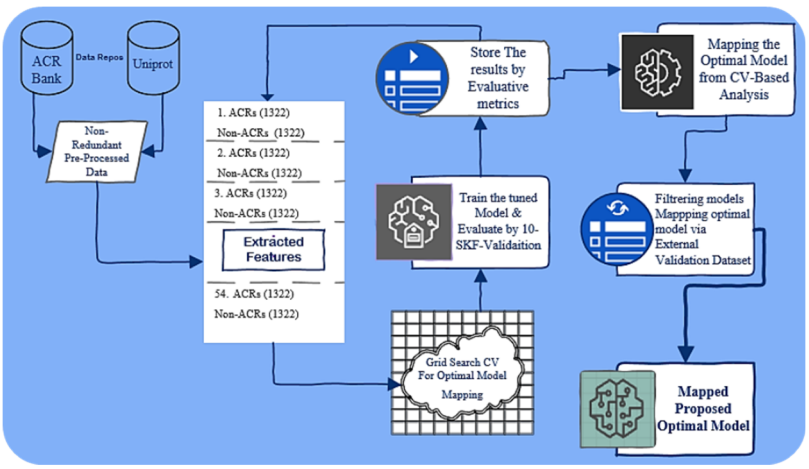

Fig. 1. Generic Framework of how we have mapped out the optimal ML-model for ACR-Predictor

Results: Using 10-fold cross validation we have achieved an area under the precision-recall curve of 0.96 using SVM and sequence derived structural descriptors. Similarly, using external validation dataset we have also achieved an area under the precision-recall curve of 0.97. These results show that our proposed predictive model predicts Anti-CRISPR (ACR) proteins with high precision (all the predicted ACRs are actually ACRs) and high recall (all the actual ACRs predicted correctly as ACRs). We have also checked the generalization performance through GO term enrichment analysis where we found a significant overlap between the terms of predicted and actual ACR proteins.
Conclusion: Thus, on the basis of deduced results, a reliable and appropriate SVM-based approach is carved out, based on selected optimal features to identify the candidate ACR's precisely and to deal with experimentally restrictive time and cost trade-offs.

References

1. Carlaw T.M., Zhang L.H., Ross C.J.D. CRISPR/Cas9 Editing: Sparking Discussion on Safety in Light of the Need for New Therapeutics. *Hum Gene Ther*. 2020;31(15-16):794-807. doi: 10.1089/hum.2020.111.