[BIOINFO 2023]

Sequence-based prediction of anti-CRISPR proteins using transformer model

Chan-Seok Jeong[1,*]

[1] *Biomedical Informatics Team, Korea Institute of Science and Technology Information*
**Corresponding author: jcs@kisti.re.kr*

The emergence of Anti-CRISPR, a natural inhibitor of the CRISPR-Cas system within prokaryotic immune systems, offers a valuable tool for preparing for post-translational regulation of the CRISPR-Cas system and mitigating the potential side effects of gene editing technology. While experimental approaches have been developed to identify anti-CRISPR proteins, the utilization of bioinformatic prediction holds promise for a more cost-effective screening strategy. However, the development of algorithms is hindered by the scarcity of verified anti-CRISPR data and the constraints of sequence similarity. In this study, we introduce a novel approach that leverages pre-trained protein language models for sequence-based anti-CRISPR prediction. This method integrates a Transformer-based sequence encoder with classification layers, and it is further fine-tuned for validated anti-CRISPR data. Our approach outperforms existing predictors that rely on ensemble boosting algorithms and achieves 2.1 times higher sensitivity on independent datasets at 95% specificity. In addition, attention structure analysis reveals the model's capacity to recognize critical residues related to anti-CRISPR functionality, even in the absence of explicit learning on these regions. Consequently, the characteristics of being able to predict using only amino acid sequences and providing important residue information make it suitable for large-scale data analysis and functional studies.