preliminary result 1: base crispAI on PRIDICT 90k dataset

Peiheng Lu

November 15, 2024

To start testing the viability of crispAI model on the prime editing dataset, the moderately sized PRIDICT 90k dataset was modified to fit the input format of the crispAI model. At the first stage, the model itself underwent very little modification, only taking the sgRNA portion of the prime editors as input.

The crispAI model requires the 73bp long flank sequence of the target site, which would be used to calculate the moving window GC content, NuPoP affinity and occupancy, as well as BDM score[1]. However, the wide target sequence in the PRIDICT 90k dataset was only 99bp long. As a result, a pipeline was developed to map each target site to the human genome reference sequence, and extract the 73bp flank sequence from the reference genome using the toolkit provided by crispAI authors.

PRIDICT provided the HGVS(Human Genome Variation Society) notation of the intended edits, as well as the direction of the target strand(forward or reverse, in relation to the reference genome).

References

[1] Florian Störtz, Jeffrey K. Mak, and Peter Minary. "piCRISPR: Physically Informed Deep Learning Models for CRISPR/Cas9 off-Target Cleavage Prediction". In: *Artificial Intelligence in the Life Sciences* 3 (Dec. 2023), p. 100075. ISSN: 26673185. DOI: 10.1016/j.ailsci.2023.100075. URL: https://linkinghub.elsevier.com/retrieve/pii/S2667318523000193 (visited on 10/14/2024).