DNA-based MSA Transformer: Advancing DNA Feature Prediction

Sukhwan Park[1], Martin Steinegger[1,2,3,4,*]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, South Korea
[2]School of Biological Sciences, Seoul National University, Seoul 08826, South Korea
[3]Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, South Korea
[4]Artificial Intelligence Institute, Seoul National University, Seoul 08826, South Korea
*Corresponding author: martin.steinegger@snu.ac.kr

In recent years, the advent of transformer-based language models have markedly changed the bioinformatics landscape, delivering state-of-the-art performance in diverse tasks ranging from protein structure prediction to DNA feature identification. Nevertheless, a significant gap persists in current DNA language models: the underutilization of homology signals derived from Multiple Sequence Alignments (MSAs). To harness the rich evolutionary information of whole-genome MSAs, we have trained a novel DNA-based MSA Transformer (DNA-MSAT). This model was trained using clustered whole-genome MSAs of 100 vertebrate species including the human genome as reference. We demonstrate that embeddings generated by DNA-MSAT have higher discriminative power on DNA features (e.g. transcription factor binding sites, silencers, promoters, etc.) relative to the embeddings of the widely used DNABERT model. After fine-tuning, DNA-MSAT attained a coding region prediction accuracy of 0.976, outperforming the twenty times larger Nucleotide Transformer, which scored 0.952. We are in the process of extending DNA-MSAT to many other DNA annotation tasks and plan to deliver a comprehensive free and open-source platform to improve our understanding of not only well-studied model species, but also the remainder of the tree-of-life.