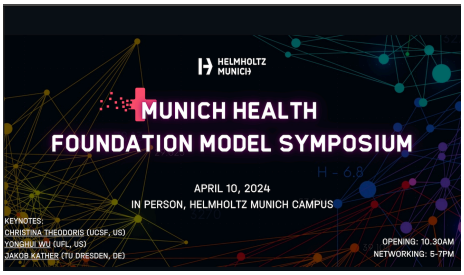

Contribution ID: **2** Type: **Talk**

Examining the effectiveness of foundation models on human 3'UTR sequences

Foundation models like DNABERT and Nucleotide Transformer have recently gained a lot of popularity in the field of DNA research. Adopted from Natural Language Processing, these models are trained in a self-supervised manner on vast amounts of genomic data. Once trained, foundation models offer applications for various downstream tasks, including promoter and enhancer prediction, prediction of epigenetic markers and splice sites, functional variant prioritization. However, genomic language models are typically trained and evaluated on entire genomes, ignoring genome partitioning into distinct functional regions.

In our work, we develop a set of 3'UTR-specific tasks to study the performance of language models on human 3'UTR sequences. These tasks include identification of binding motifs of RNA binding proteins, detection of functional genetic variants, prediction of expression levels in massively parallel reporter assays, and estimation of mRNA half-life. In total, we test three established genome-wide foundation models as well as five transformer models that we specifically train on 3'UTR sequences from 241 mammalian species.

We demonstrate that the models specifically trained on 3'UTR sequences exhibit superior performance in three out of four downstream tasks compared to their genome-wide counterparts. These findings emphasize the significance of accounting for genome partitioning into distinct functional regions while training and evaluating foundation models. We also note that the proposed set of 3'UTR-specific tasks may serve as a benchmark for assessing the performance of future models.

The results of our work are currently available as a bioRxiv preprint: https://www.biorxiv.org/content/10.1101/2024.02.09.579631v1

Primary authors: Dr HEINIG, Matthias (Helmholtz Zentrum München, Institute of Computational Biology); Dr VILOV, Sergey (Helmholtz Zentrum München, Institute of Computational Biology)

Session Classification: Break + Posters session

Track Classification: Poster