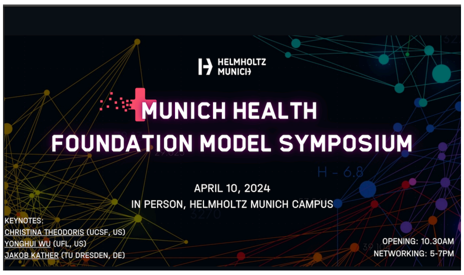
HELMHOLTZ
MUNICH
MUNICH HEALTH
FOUNDATION MODEL SYMPOSIUM
APRIL 10, 2024
IN PERSON, HELMHOLTZ MUNICH CAMPUS
KEYNOTES:
CHRISTINA THEODORIS (UCSF, US)
YONGHUI WU (UFL, US)
JAKOB KATHER (TU DRESDEN, DE)
OPENING: 10.30AM
NETWORKING: 5-7PM

Contribution ID: **31** Type: **Talk**

DNA language models provide a map of nucleotide and cis-regulatory code interactions in the genome

DNA and RNA are made up of contiguous chains of nucleotides, yet interactions between nucleotides are not strictly linear. Regions inside these sequences interact to form structures, define splicing, and establish cis-regulatory codes, among other genomic functions. DNA language models (DNA LMs) have been shown to capture local motifs and differentiate between regions in the genome with distinct functions such as untranslated regions, protein coding regions, introns and intergenic areas. However, the ability of DNA LMs to capture long-range nucleotide and cis-regulatory code interactions remains unexplored.

Here we show that DNA LMs provide a detailed map of long-range interactions in the genome, offering a novel tool for their discovery and analysis. We evaluate a DNA LM trained on multiple species, showing its ability to capture RNA secondary structures and splicing. Furthermore we identify and characterize newly found long-range interaction patterns, including potential novel RNA secondary structures that require further experimental study. Our analysis reveals that most interactions follow a power-law relationship with distance, with parameters varying across genomes. We furthermore investigate the interaction maps from the DNA LM generated for artificially manipulated sequences, showcasing its ability to predict repeat regions without memorizing them and to understand base pairing and splicing rules.

Overall, this study demonstrates the ability of DNA LMs to reveal complex long-range genomic interactions while presenting a straightforward and universally applicable method to do so for any DNA LM designed to predict nucleotide probabilities.

Primary authors: TOMAZ DA SILVA, Pedro (TUM); Mr KAROLLUS, Alexander (TUM); Mr HINGERL, Johannes (TUM); Dr HERNANDEZ-ALIAS, Xavier (TUM); Mr WAGNER, Nils (TUM); Prof. GAGNEUR, Julien (TUM)

Session Classification: Break + Posters session

Track Classification: Poster