Classifying SARS-CoV-2 genomes using advanced Natural Language Processing

Pushkar Ambastha • Navin Patwari • Jaskirat Singh • Shubhanshu Sahu

BACKGROUND

Rapid genomics progress enables decoding viral genetic codes, which is pivotal for grasping evolution, pathogenesis, and therapies. Though robust, traditional sequence alignment and phylogenetics used for virus classification may struggle with vast genomic data. This drives interest in NLP techniques to unravel viral genetic intricacies and complement bioinformatics, informing therapeutics and outbreak preparedness.

PRINCIPLE

The amalgamation of diverse models in our approach enhances viral genome classification precision. Probabilities from these models are crucial indicators for assigning cDNA sequences to distinct viral classifications.

METHODS

We primarily applied classical machine learning methods for sequence classification, including random forests and SVCs. We developed a transformer pipeline involving:

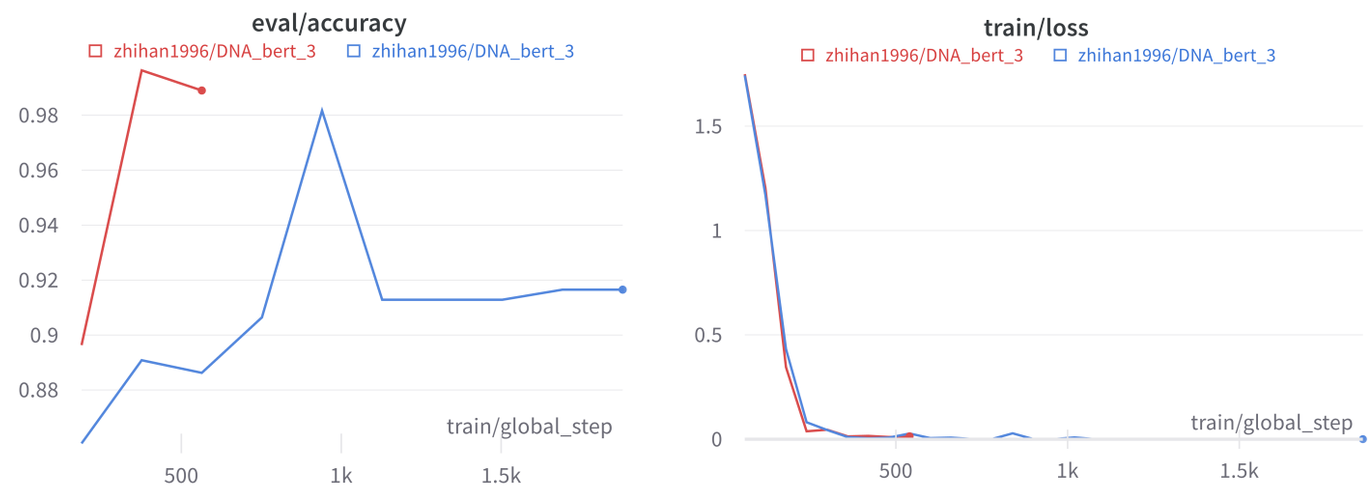
- Tokenization
- Pre-Training
- SARS-CoV dataset Fine-tuning

to optimize performance and resources. The transformer leverages self-attention to model contextual relationships between input sequence tokens.

RESULTS

Model Name	Accuracy	Modality
DNA BERT	0.991	Specifically for DNA sequence
Generic Entity Recognition	0.985	Specifically for gene data
Multinomial Naive Bias	0.072	ML Based
Random Forest	0.071	ML Based
K-Nearest Neighbors	0.071	ML Based
LinearSVC	0.042	ML Based
MiniLM-L12-H384-uncased	0.018	Original Language models

Table: Model performance and modalities



Above: evaluation accuracy (F1), training loss depicted with increase in global training steps

We developed DNABERT, a novel pre-trained bidirectional encoder for capturing global, transferable understanding of DNA sequences using nucleotide contexts. Compared to widely used regulatory element predictors, this single model achieves state-of-the-art performance in predicting promoters, splice sites, and transcription factor binding sites after minimal fine-tuning with small labeled data. We anticipate DNABERT can be readily fine-tuned for diverse sequence analysis tasks.

APPLICATIONS

The complex gene regulatory code, with its challenges in conventional informatics methods, is especially pronounced in scenarios with limited data. This has wide implications for precision medicine, drug development, and bioinformatics tools.