Exploring the Embedding Methods in Genomic Language Models

Anisa Habib
University of Utah

UUCS-24-005

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

26 April 2024

Abstract

Language Models (LMs) have revolutionized natural language processing, excelling in language translation, sentiment analysis, and text generation. Researchers have proposed using LMs to learn generalizable features from DNA, aiming to fine-tune these models for diverse prediction tasks. While several LMs trained on DNA sequences now exist, they vary in tokenization methods, the types and amounts of data used for training, and the specific tasks they are fine-tuned for. Existing benchmark reports often lack comprehensive coverage and consistency in reported metrics. To address this gap and explore the impact of different encoding schemes for DNA, this study conducts benchmarking tests on standard tasks to assess and compare existing models' performance capabilities. Additionally, we construct our own fine-tuning task to perform preliminary investigations on whether an LM can accurately identify the locations of prophage sequences integrated into the bacterial genome. Our findings suggest that model accuracy varies depending on the task, with no single model performing best across all tasks. We observed that tasks exhibit different levels of difficulty, and there is a wide distribution of variation in performance even with the same model.