Decoding DNA: Exploring the Impact of Tokenization on Genomic Language Models

Anisa Habib, LeAnn Lindsey, Dr. Hari Sundar (Advisor), *University of Utah*

Introduction

Large Language Models have gained considerable popularity over the past years, owing to their capacity to be trained on unlabeled data and extract meaningful insights from human language.

Recent models such as Nucleotide Transformer, DNABERT, and HyenaDNA are trained on DNA to complete a variety of genomic tasks. However, these models are all:

a. trained with different tokenization methods
b. trained on different types and amounts of data
c. finetuned on different tasks

In other words, each model was built using different data representations, the amount of information captured per token varying.

To investigate the impact of different encoding schemes for DNA sequences, we ran benchmarking tests on standard tasks using existing models to determine and compare their performance capabilities.

Background

Why does tokenization matter? Tokenization can increase the total information in a given context window.

Figure 1 – Central Dogma (Ngyuen et. al 2024)

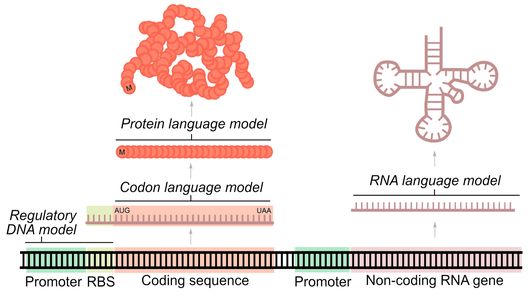


Figure 2 – Tokenization Methods

Nucleotide (*HyenaDNA*)	A T G T T C A G G C C G A C
Codon (*GenSLM*)	ATG TTC CAG GCC GAC
kmer (*DNABERT, NT*)	ATGGTT TGGTTC GGTTTG GTTTGG TTTGGT
BPE (*DNABERT2*)	AT GGTT CAGGC CGGA GCC

Figure 3 – Bacteria vs. DNABERT2 BPE Vocabularies

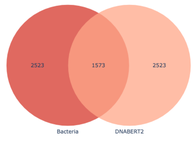
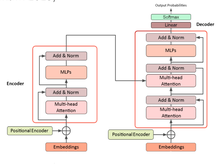


Figure 4 - The Transformers Architecture (Nyandwi 2023)



Methods

1. Download models and retrieve benchmark datasets, locations provided by authors.
   a. An 8:1:1 training, development, test split was used for all datasets.
2. Run each model on every task 10 times each.
   a. All models finetuned using the same set of parameters provided.
   b. All tasks run on 1 NVIDIA Tesla V100 GPU provided by the PSC's Bridges-2.
3. Report the results over the Matthew Correlation Coefficient.

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

A more reliable statistic than accuracy or F1 score in binary classification evaluation.

- 1 = best results in the four confusion matrix categories (true positives, false negatives, true negatives, and false positives)
- 0 = no agreement; prediction is random

Results

Figure 5 – Example Task Results by Model

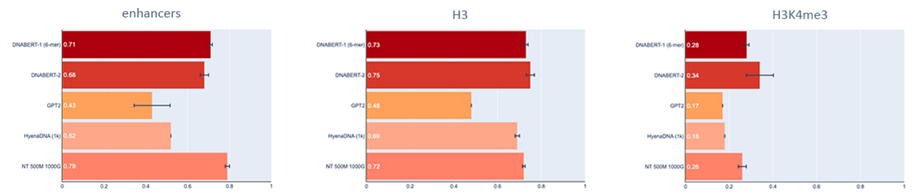


Figure 6 – Mean MCC Benchmark Results by Model

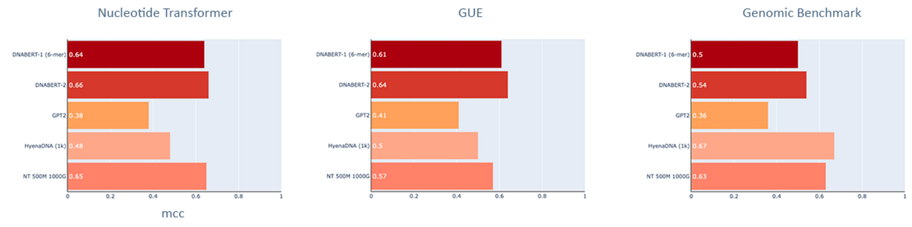
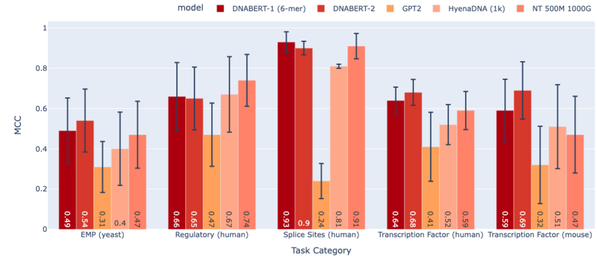


Figure 7 - Mean MCC Task Category Results by Model



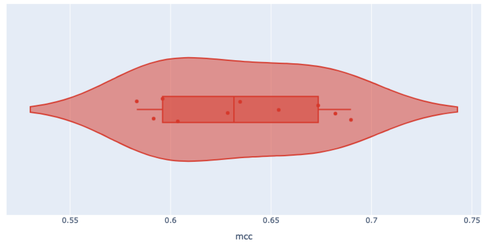
Example Task: 'enhancers'

- Regulatory elements are regions of DNA that play a crucial role in controlling gene expression.
- Predict if an enhancer region is present in a DNA sequence.

Table 1 - Tested Language Models

Model	Parameters	Tokenization	Finetune Benchmark	Context Window	PreTraining Data	PreTraining Time	PreTraining Hardware
GPT-2	124M - 1.5B	BPE		1024	WebText	4 days - 1 month	256 Tesla P100
DNABERT-1	117 M	kmer	GUE	512	Human Genome	25 days	8 Nvidia 2080Ti
DNABERT-2	117 M	BPE	GUE	1000	Human Genome + 135 Other Species	14 days	8 Nvidia 2080Ti
Hyena-DNA	0.5M - 6.6M	Nucleotide	Genomic Benchmark	1000 - 1M	Human Genome	80 min - 4 weeks	1 Nvidia A100
Nucleotide Transformer	500M - 2.5B	kmer	Nucleotide Transformer	1000	3202 Human Genomes + 850 Other Species	14 days	16 x 8 Nvidia A100 (128 Total)

Figure 8 - DNABERT-2 Promoter 300 TATA MCC Variation Across Replications



Discussion

With differences in training data and tokenization, model accuracy appears to vary depending on the task. It is not clear from these experiments that one model performs the best on all tasks. We observe that:

a. Different tasks have different degrees of difficulty.
b. There is a wide distribution of variation with replication, even with the same model.
c. DNABERT-2 appears to perform well on tasks with longer input sequences.
d. NT appears to perform well on tasks with shorter input sequences.

Future Work

More work is needed to determine why some models do well on some tasks and other models do well on other tasks. Future research may include:

a. Training a HyenaDNA model using BPE tokenization to see if that increases accuracy.
b. Training DNABERT-2 on just the Human Genome to see what portion of the accuracy on different tasks is due to tokenization vs. multi-species data.

References

1. Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7
2. Hugo Dalla-Torre *et al.*, "The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics," *bioRxiv*, p. 2023.01.11.523679, Jan. 2023, doi: 10.1101/2023.01.11.523679.
3. Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, Aug. 2021, doi: 10.1093/bioinformatics/btab083.
4. E. Nguyen *et al.*, "HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution." arXiv, Nov. 14, 2023. doi: 10.48550/arXiv.2306.15794.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.
6. J. Nyandwi, "The Transformer Blueprint: A Holistic Guide to the Transformer Neural Network Architecture." GitHub, July 29. 2023 https://deeprevision.github.io/posts/001-transformer/
7. Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome." arXiv, Jun. 26, 2023. doi: 10.48550/arXiv.2306.15006.