22 May

Pedro Pinto

Vrije Universiteit Amsterdam

HyenaDNA and DNABert 2

The benchmarks were run using frozen model embeddings and a simple logistic regression classifier, without any model fine-tuning. Each benchmark was repeated over 5 runs with different random seeds to ensure consistency. Initially, HyenaDNA was implemented first, which is why the style of its code notebooks differs from the later DNABERT-2 benchmarks. I will normalize the style and remove the inconsistencies in the coming days, while I am considering fine tuning.

Below is a summary of average performance across benchmarks:

Benchmark	Model	Accuracy	F1	ROC AUC
Pathogenicity	HyenaDNA	0.5904	0.6079	0.6202
	DNABERT-2	0.5575	0.5608	0.5819
TFBS	HyenaDNA	0.5843	0.6389	0.6249
	DNABERT-2	0.5855	0.6079	0.6167
Shuffled vs. Real DNA	HyenaDNA	0.6950	0.6659	0.7571
	DNABERT-2	0.9261	0.9260	0.9772

Both models perform moderately on pathogenicity classification and TFBS classification tasks using raw sequence alone, with HyenaDNA slightly outperforming DNABERT-2 in those settings. However, in the shuffled DNA benchmark, DNABERT-2 achieves near-perfect discrimination, indicating it has strongly internalized DNA sequence structure during pre-training.

These results suggest that while pretrained embeddings can separate biologically meaningful signals to some extent, there is room for improvement.