HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution

Eric Nguyen[*,1], Michael Poli[*,1], Marjan Faizi[2,*],
Armin W. Thomas[1], Callum Birch Sykes[3], Michael Wornow[1], Aman Patel[1],
Clayton Rabideau[3], Stefano Massaroli[4], Yoshua Bengio[4], Stefano Ermon[1],
Stephen A. Baccus[1,†], Christopher Ré[1,†]

November 15, 2023

Code available at https://github.com/HazyResearch/hyena-dna

presented by: Felix Dyck - felix.dyck@uni-bielefeld.de

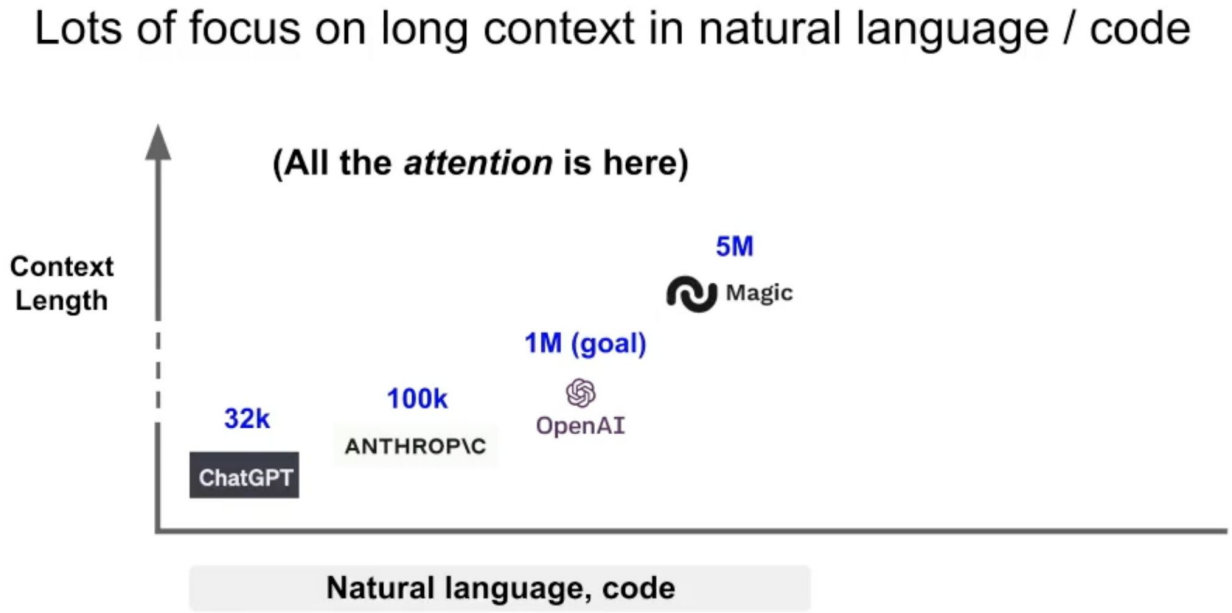
05.12.24

Content

- Motivation
- Background
- HyenaDNA Block
- Experiments
- Sequence length warm-up
- Task Adaptation

Long Context

Human Genome has 3.2B nucleotides



Background

Nucleotides in a DNA sequence ACGTACGTCGTACGTC…

Previous work:

- typically context length of 512 - 4k tokens
- resolution: usually tokens are not on “nucleotides-level” -> “K-mers” (3-5 nucleotides)

Problems:

- long context is computationally expensive
- “SNPs” (Single-Nucleotide Polymorphisms): variation at a single nucleotide in the DNA sequence

Perplexity vs Context

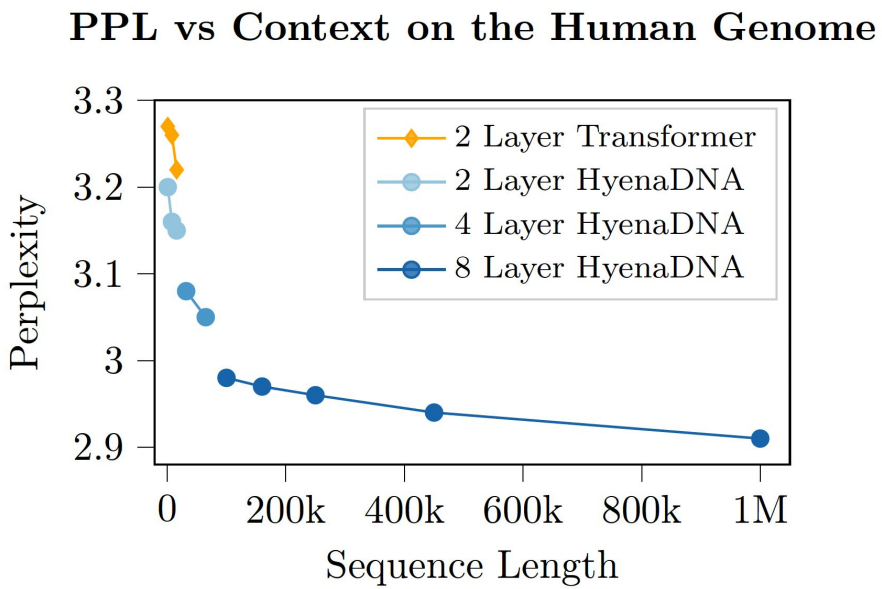


Figure 1.2: Pretraining on the human reference genome using longer sequences leads to better perplexity (improved prediction of next token).

HyenaDNA Block Architecture

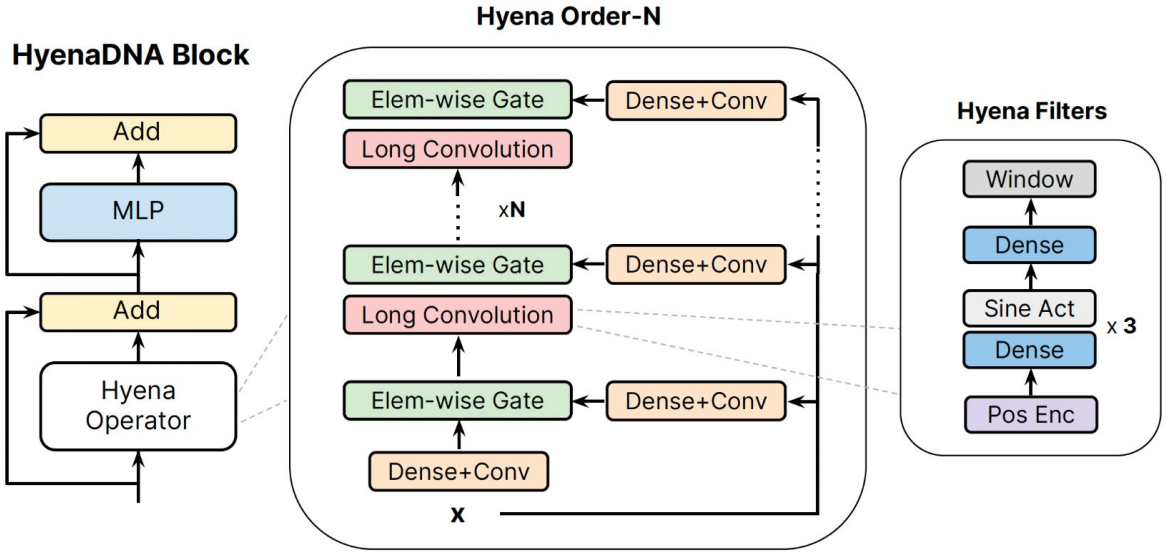


Figure 1.3: HyenaDNA block architecture. A Hyena operator is composed of long convolutions and element-wise gate layers. The gates are fed projections of the input using dense layers and short convolutions. The long convolutions are parameterized *implicitly* via an MLP that produces the convolutional filters. The convolution itself is evaluated using a Fast Fourier Transform convolution with time complexity $\mathcal{O}(L \log_2 L)$.

Runtime

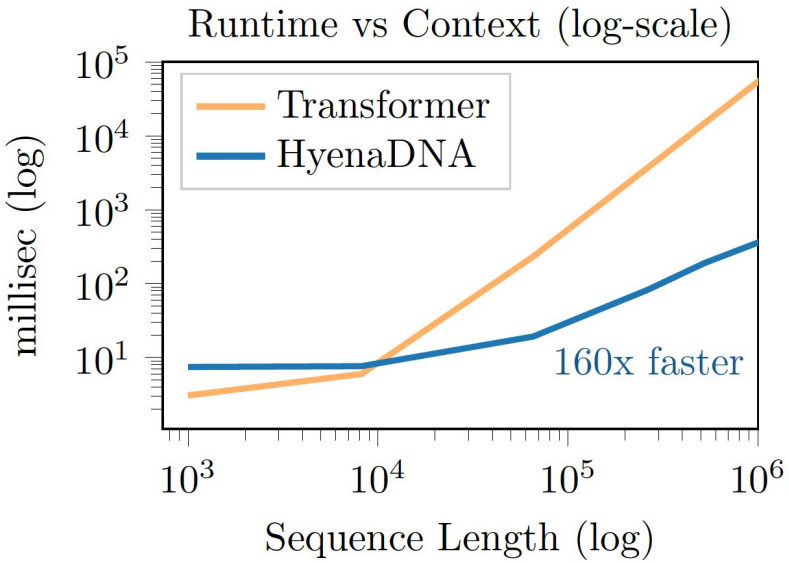


Figure 4.1: Runtime (forward & backward pass) for Transformer and HyenaDNA: 2 layers, width=128, gradient checkpointing, batch size=1, A100 80GB. At 1M tokens HyenaDNA is **160x faster** than Transformer.

Single Nucleotide Resolution

Table 4.1: **GenomicBenchmarks** Top-1 accuracy (%) for pretrained HyenaDNA, DNABERT and Transformer (GPT from 4.1), and the previous SotA baseline CNN (scratch).

Dataset	CNN	DNABERT	GPT	HyenaDNA
Mouse Enhancers	69.0	66.9	80.1	**85.1**
Coding vs Intergenomic	87.6	**92.5**	88.8	91.3
Human vs Worm	93.0	96.5	95.6	**96.6**
Human Enhancers Cohn	69.5	74.0	70.5	**74.2**
Human Enhancers Ensembl	68.9	85.7	83.5	**89.2**
Human Regulatory	93.3	88.1	91.5	**93.8**
Human Nontata Promoters	84.6	85.6	87.7	**96.6**
Human OCR Ensembl	68.0	75.1	73.0	**80.9**

Ultralong-Range Genomics

- single-nucleotide polymorphisms (SNPs)
- quantifying the functional effects of non-coding variants

Table 4.3: **Chromatin profile prediction** Median AUROC computed over three categories: Transcription factor binding profiles (TF), DNase I-hypersensitive sites (DHS) and histone marks (HM).

Model	Params	Len	AUROC		
			TF	DHS	HM
DeepSEA	40 M	1k	95.8	92.3	85.6
BigBird	110 M	8k	96.1	92.1	88.7
HyenaDNA	7 M	1k	**96.4**	**93.0**	86.3
	3.5 M	8k	95.5	91.7	**89.3**

Species classification

- DNA sequences from 5 different species
- struggle on shorter sequences of length 1024

Table 4.5: **Species classification** Top-1 accuracy (%) for 5-way classification (human, lemur, mouse, pig, hippo). The ✗ symbol indicates infeasible training time.

Model	Len	Acc
Transformer	1k	55.4
HyenaDNA	1k	61.1
Transformer	32k	88.9
HyenaDNA	32k	93.4
Transformer	250k	✗
HyenaDNA	250k	97.9
Transformer	450k	✗
HyenaDNA	450k	99.4
Transformer	1M	✗
HyenaDNA	1M	**99.5**

Sequence length warm-up

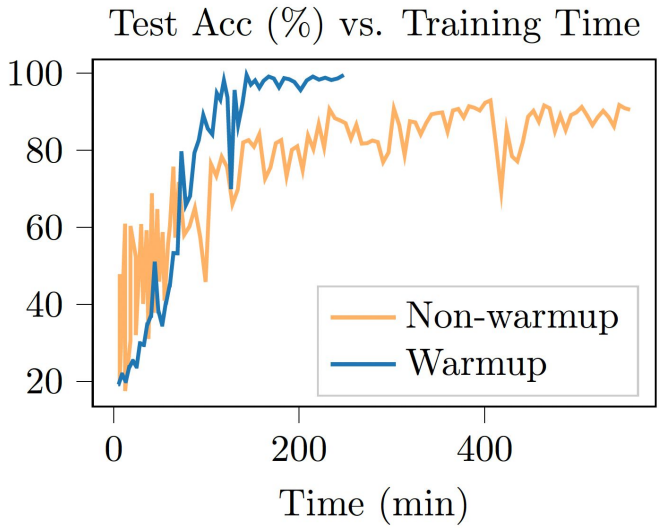


Figure 3.2: Sequence length warm-up reduces the training time of HyenaDNA at sequence length 450k by 40% and boosts accuracy by 7.5 points on species classification.

Task Adaptation with Soft Prompt Tokens

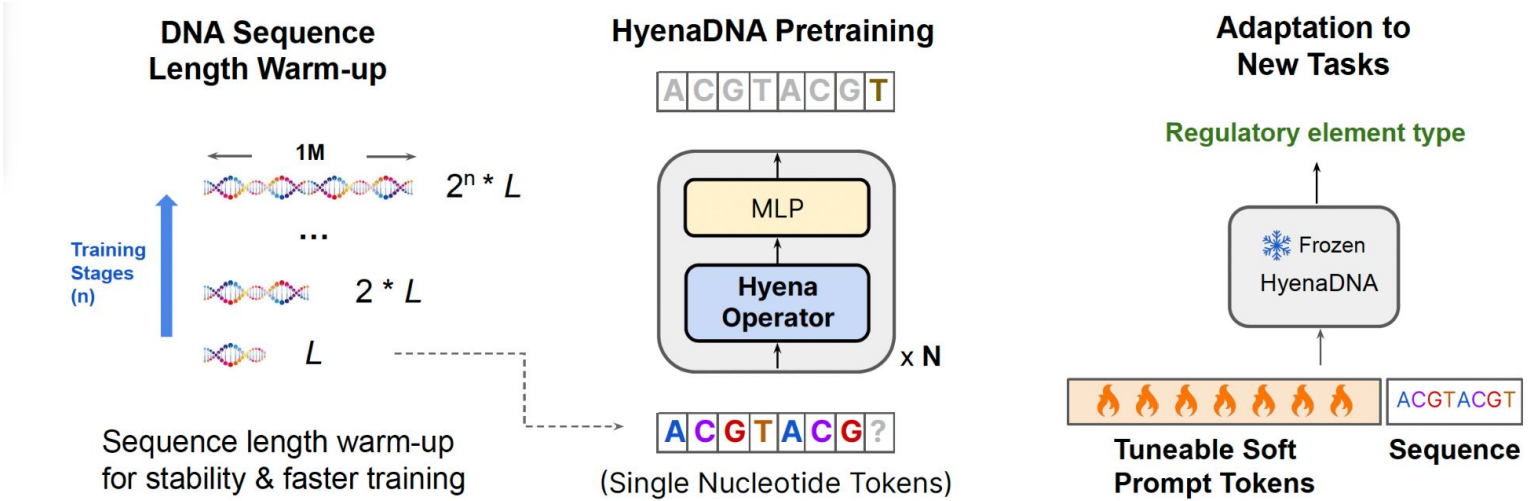


Figure 1.1: HyenaDNA recipe for long-range foundation models in genomics. The HyenaDNA architecture is a simple stack of Hyena operators (Poli et al., 2023) trained using next token prediction. (See Fig. 1.3 for block diagram of architecture). We introduce a new sequence length scheduling technique to stabilize training, and provide a method to leverage the longer context length to adapt to novel tasks without standard fine-tuning by filling the context window with learnable soft prompt tokens.

Fine Tuning HyenaDNA

Tasks:

- Identify Enhancer Regions in a DNA Sequence
- Species classification

Receipt:

- Pretrained Model: HyenaDNA
- Input DNA Sequence
- Soft Prompt Tokens
- Fine-Tuning:
  - Soft Prompt tokens are optimized for the task
- Output: Probability scores for:
  - each nucleotide being part of an enhancer
  - dna being each species

Example:

Input: ACGTACGTCGTACGTC…

-> [0.1, 0.2, 0.3, 0.4, ...]

- Soft Prompt: [P1, P2, P3]

-> P1 = [0.01, -0.05, 0.07, …], P2 = [0.1, 0.03, -0.02, …], P3 = [0.05, 0.07, -0.01, ...]

- Combined Input: [P1, P2, P3, 0.1, 0.2, 0.3, 0.4, …]

- Predictions: [0.01, 0.85, 0.10, 0.90, …]
(Species A, Species B, …)

Summary

- Sequence length of up to 1 million turned out to be beneficial.
- Observing single nucleotides as individual tokes is important.
- Using long convolutions is especially more efficient the longer the sequence gets
- “Warm-up” learns shorts sequences first and then gradually increases the length leading to better performance faster and in the end.

Thank you for Listening!

Hyperparameters used for Training

Table A.3: GenomicBenchmarks hyperparameters for HyenaDNA and the baseline Transformer (GPT from 4.1), which uses FlashAttention (Dao et al., 2022a).

	Transformer	HyenaDNA
Layers	2	2
Width	128	128
Parameters	529k	436k
Learning rate	1-6e^{-4}	1-6e^{-4}
Weight decay (model)	0-0.2	0-0.2
Weight decay (Hyena layers)	-	0
Embed dropout	0-0.2	0.0-0.3
Resid dropout	0-0.2	0-0.3
Num heads	8	-
Optimizer	AdamW	
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
LR scheduler	Cosine decay	
Batch size	128-1024	
Training epoch	100	
Reverse complement aug.	true/false	
Sequence lengths	200-4800	

Table 1. Genomic Benchmarks. Top-1 accuracy (↑) across 5-fold cross-validation (CV) for pretrained HyenaDNA, Mamba NTP, Caduceus models, and a supervised CNN baseline (trained from scratch). Best values per task are **bolded**, second best are *italicized*. Error bars indicate the difference between the maximum and minimum values across 5 random seeds used for CV.

	CNN (264K)	HyenaDNA (436K)	Mamba (468K)	Caduceus w/o Equiv. (470K)	Caduceus-Ph (470K)	Caduceus-PS (470K)
Mouse Enhancers	0.715 ±0.087	*0.780* ±0.025	0.743 ±0.054	0.770 ±0.058	0.754 ±0.074	**0.793** ±0.058
Coding vs. Intergenomic	0.892 ±0.008	0.904 ±0.005	0.904 ±0.004	0.908 ±0.003	**0.915** ±0.003	*0.910* ±0.003
Human vs. Worm	0.942 ±0.002	0.964 ±0.002	0.967 ±0.002	*0.970* ±0.003	**0.973** ±0.001	0.968 ±0.002
Human Enhancers Cohn	0.702 ±0.021	0.729 ±0.014	0.732 ±0.029	0.741 ±0.008	**0.747** ±0.004	*0.745* ±0.007
Human Enhancer Ensembl	0.744 ±0.122	0.849 ±0.006	0.862 ±0.008	0.883 ±0.002	*0.893* ±0.008	**0.900** ±0.006
Human Regulatory	0.872 ±0.005	0.869 ±0.012	0.814 ±0.211	0.871 ±0.007	*0.872* ±0.011	**0.873** ±0.007
Human OCR Ensembl	0.698 ±0.013	0.783 ±0.007	0.815 ±0.002	0.818 ±0.003	**0.828** ±0.006	*0.818* ±0.006
Human NonTATA Promoters	0.861 ±0.009	0.944 ±0.002	0.933 ±0.007	0.933 ±0.006	**0.946** ±0.007	*0.945* ±0.010

Table A.4: **GenomicBenchmarks Top-1 accuracy (%)** GPT is the causal Transformer from 4.1, HyenaDNA k-mer uses a 6-mer tokenizer, and HyenaDNA bidirection is a bidirectional version of the Hyena operator.

Model	GPT	GPT	HyenaDNA	HyenaDNA	HyenaDNA k-mer	HyenaDNA bidirection	DNABERT
Pretrained	no	yes	no	yes	no	no	yes
Mouse Enhancers	79.3	79.3	84.7	**85.1**	81.8	80.6	66.9
Coding vs Intergenomic	89.3	91.2	90.9	91.3	86.7	90.3	**92.5**
Human vs Worm	94.8	**96.6**	96.4	**96.6**	92.9	95.9	96.5
Human Enhancers Cohn	67.7	72.9	72.9	**74.2**	69.8	72.1	74.0
Human Enhancers Ensembl	79.0	88.3	85.7	**89.2**	88.0	85.9	85.7
Human Regulatory	90.2	91.8	90.4	**93.8**	90.2	89.1	88.1
Human Nontata Promoters	85.2	90.1	93.3	**96.6**	83.5	88.5	85.6
Human OCR Ensembl	68.3	79.9	78.8	**80.9**	70.2	75.3	75.1

Some more References…

Video-presentation of the paper:

https://youtu.be/haSkAC1fPX0?si=v28n75mdSjTURvay

Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling:

https://arxiv.org/pdf/2403.03234

Articles:

- https://hazyresearch.stanford.edu/blog/2023-06-29-hyena-dna
- https://aibusiness.com/ml/hyenadna-a-large-language-model-trained-on-human-genome-sequences