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

by Eric Nguyen, Michael Poli, Marjan Faizi

Image source: https://africafreak.com/spotted-hyena-facts (last visited: 02.09.2024)

PRESENTATION BY FLORIAN DRÖSSLER

Overview

- Background Information
- Motivation
- Architecture
- Experiments with the model
- Conclusion

Background in Genomics

- DNA sequences carry genetic instructions
- Sequences consist of chains of nucleotides, represented by the letters A, T, C, and G
- Understanding relationships between sequences and biological functions is key to advancements
- BUT: The human genome is about 3.2 billion nucleotides long making it extremely complex
- DNA contains long-range dependencies, interaction between distant parts can influence gene regulation

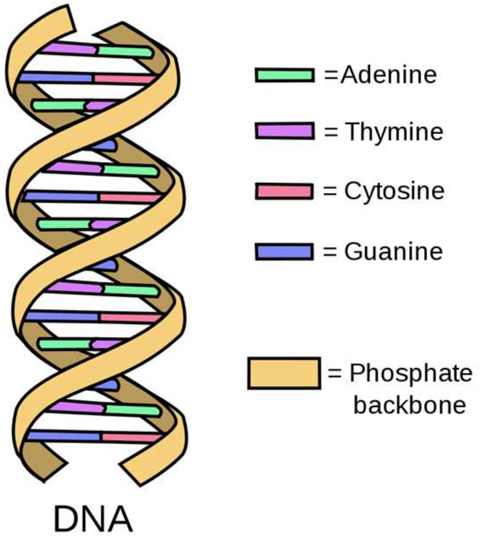


Image source: https://www.ashg.org/discover-genetics/building-blocks/ (last visited: 02.09.2024)

Convolutional vs Transformer Models

Convolutional:

- Convolutional layers apply filter to local regions
- Detects local patterns
- Filters share parameters (weights)
- Generally difficult to capture global context
- Often used in Image processing

Transformer:

- Use attention mechanism to process entire input simultaneously
- Capture global dependencies
- No parameter sharing -> more flexible
- Designed to capture long-range dependencies and global context
- Computational intense
- Used in NLP and Genomics

Challenges in Genomic Modeling

- Traditional genomic models, struggle processing long DNA sequences
  - Limits context length to a few thousand tokens at most (typically 512 to 4,096 tokens), < 0.001% of the human genome
  - Difficult to model long-range dependencies
- Existing models rely on tokenizers or fixed k-mers (small overlapping sequences of nucleotides)
  - Helps reducing complexity but sacrifices single nucleotide resolution

HyenaDNA - Motivation

- On language H3 and Hyena achieve State of the Art (SotA) performances by stacking long convolutional layers
- HyenaDNA designed to overcome these challenges by enabling long-range genomic sequence modeling at single nucleotide resolution
- The model aims to extend the capabilities of genomic foundation models, enabling more accurate predictions and analyses in genomics

Model Structure

- HyenaDNA is a decoder-only sequence-to-sequence model
- Core component of the model is the **Hyena operator**, which replaces the traditional attention mechanism with a convolution-based approach
- Each HyenaDNA block consists of a **Hyena operator** followed by a feed-forward neural network (MLP). The operator includes:
  - **Long Convolutions**: parameterized to operate over long sequences, enabling the model to maintain context over extended stretches of DNA.
  - **Element-wise Gates**: control flow of information within model, dynamically adjusting importance of different parts of the sequence.

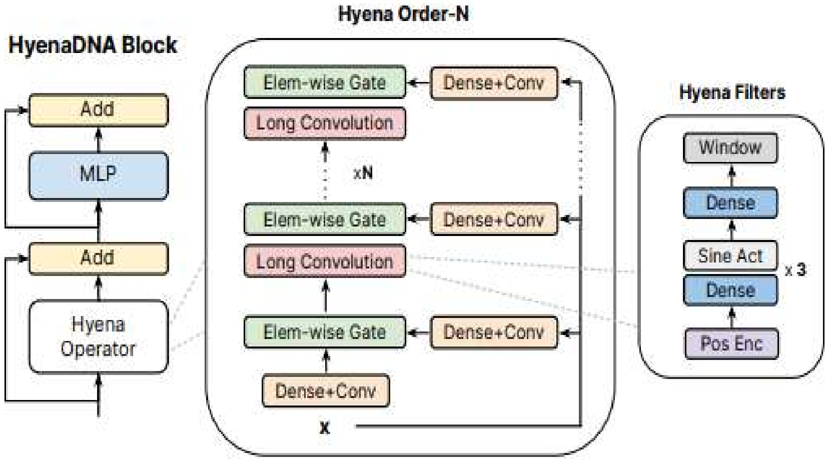


Image source: https://doi.org/10.48550/arXiv.2306.15794

Training Process

- HyenaDNA is pretrained using the human reference genome, focusing on next nucleotide prediction
- Single nucleotide tokenizer, preserving highest possible resolution, crucial for tasks where a single nucleotide difference can be significant
  - Vocabulary is minimal, including the four nucleotides plus special tokens for padding and unknown characters, ensuring the focus remains on the nucleotide-level information
- To stabilize training process with ultra-long sequences (200k+), HyenaDNA employs a sequence length warm-up technique

Training Process

- This approach allows HyenaDNA to handle sequences up to 1 million tokens effectively, making it one of the fastest and most scalable genomic models available
- At sequence lengths of 450k tokens and above, this approach has been shown to reduce training time by 40% and improve accuracy by 7.5 percentage points on species classification tasks
- Hyena Operator can be evaluated in time $\mathcal{O}(L \log_2 L)$

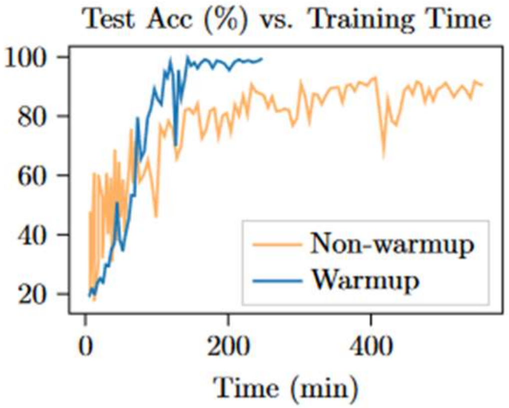


Image source: https://doi.org/10.48550/arXiv.2306.15794

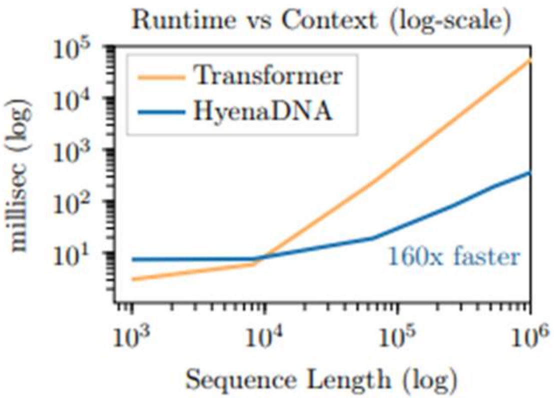


Image source: https://doi.org/10.48550/arXiv.2306.15794

Soft Prompting

- Novel adaptation technique, which involves injecting learnable tokens (as weights) directly into the input sequence
- Unlike traditional fine-tuning, soft prompting allows the model to adapt to new tasks by updating only the prompt tokens, keeping the rest of the model fixed
- This method is highly efficient, requiring fewer computational resources and less training data, making it ideal for quick adaptation to various genomic tasks

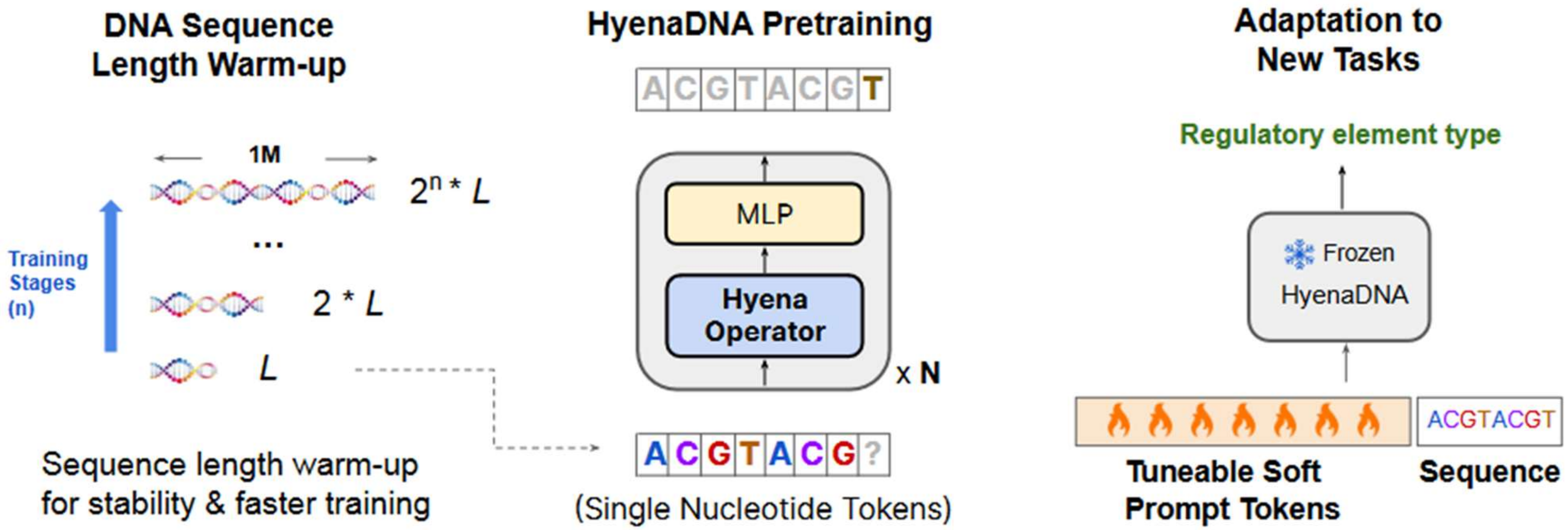


Image source: https://doi.org/10.48550/arXiv.2306.15794

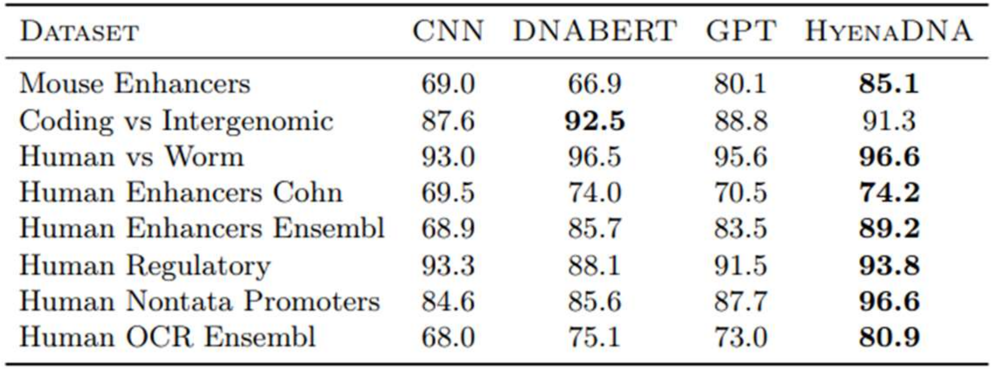
Recap

Attention-based models suffer from computational costs as sequence length grows, HyenaDNA scales efficiently, enabling the processing of sequences up to 1 million tokens long

This represents a 500x increase in context length over previous genomic models, allowing HyenaDNA to analyze large portions of the genome in a single pass

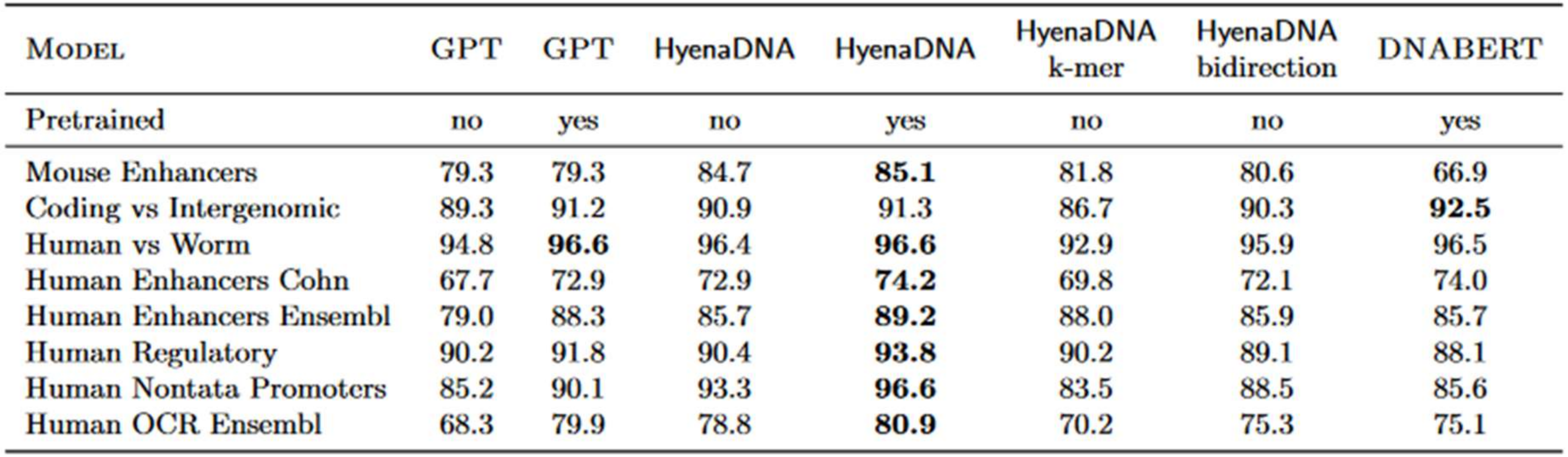
Single Nucleotide Resolution Performance

- GenomicBenchmarks Results:
  - evaluated on GenomicBenchmarks dataset, includes various tasks related to regulatory element classification and species differentiation
  - SotA results on 7 out of 8 datasets, strong performance on tasks like human enhancer identification
  - Shows HyenaDNA's ability to maintain single nucleotide resolution across entire sequences



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**

Source: https://doi.org/10.48550/arXiv.2306.15794

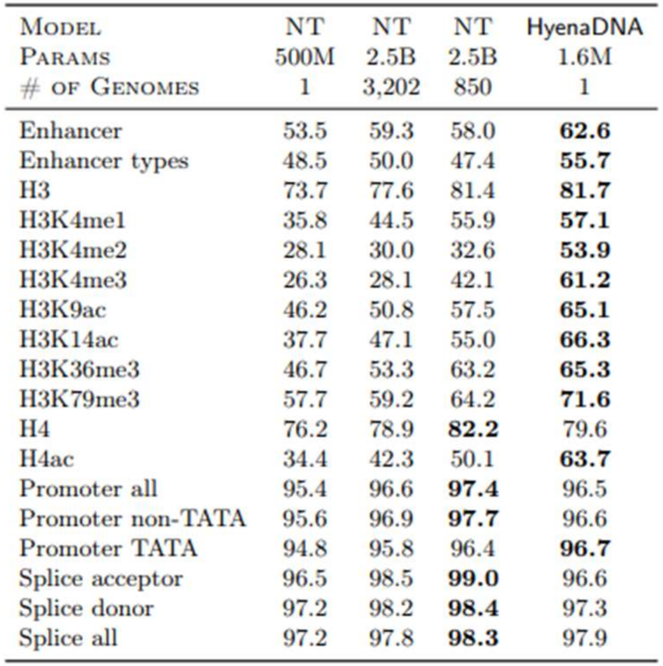


Model	GPT	GPT	HyenaDNA	HyenaDNA	HyenaDNA k-mer	HyenaDNA bidirection	DNABERT
Pretrained	no	yes	no	yes	no	no	yes
Mouse Enhancers	79.3	79.3	84.7	**85.1**	81.8	80.6	66.9
Coding vs Intergenomic	89.3	91.2	90.9	91.3	86.7	90.3	**92.5**
Human vs Worm	94.8	**96.6**	96.4	**96.6**	92.9	95.9	96.5
Human Enhancers 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

Source: https://doi.org/10.48550/arXiv.2306.15794

Single Nucleotide Resolution Performance

- Nucleotide Transformer:
  - benchmarked against the Nucleotide Transformer on 18 datasets involving tasks like enhancer and promoter identification, as well as splice site prediction
  - SotA on 12 of 18 datasets, despite using significantly fewer parameters (1.6 million compared to 2.5 billion) and much less pretraining data (one genome vs. thousands)
  - This highlights HyenaDNA's efficiency and capability to outperform larger models on complex genomic tasks



Model	NT	NT	NT	HyenaDNA
Params	500M	2.5B	2.5B	1.6M
# of Genomes	1	3,202	850	1
Enhancer	53.5	59.3	58.0	**62.6**
Enhancer types	48.5	50.0	47.4	**55.7**
H3	73.7	77.6	81.4	**81.7**
H3K4me1	35.8	44.5	55.9	**57.1**
H3K4me2	28.1	30.0	32.6	**53.9**
H3K4me3	26.3	28.1	42.1	**61.2**
H3K9ac	46.2	50.8	57.5	**65.1**
H3K14ac	37.7	47.1	55.0	**66.3**
H3K36me3	46.7	53.3	63.2	**65.3**
H3K79me3	57.7	59.2	64.2	**71.6**
H4	76.2	78.9	**82.2**	79.6
H4ac	34.4	42.3	50.1	**63.7**
Promoter all	95.4	96.6	**97.4**	96.5
Promoter non-TATA	95.6	96.9	**97.7**	96.6
Promoter TATA	94.8	95.8	96.4	**96.7**
Splice acceptor	96.5	98.5	**99.0**	96.6
Splice donor	97.2	98.2	**98.4**	97.3
Splice all	97.2	97.8	**98.3**	97.9

Source: https://doi.org/10.48550/arXiv.2306.15794

In-Context Learning Capabilities:

- Soft Prompting Performance:
  - showed that model could adapt to new tasks by simply modifying a small set of learnable prompt tokens, without full model retraining
  - As more tuneable tokens were added to the input sequences, the model's performance on novel tasks improved, reaching levels comparable to traditional fine-tuning approaches
  - Particularly useful for quickly adapting to tasks with minimal computational overhead, making it an efficient solution for dynamic genomic applications

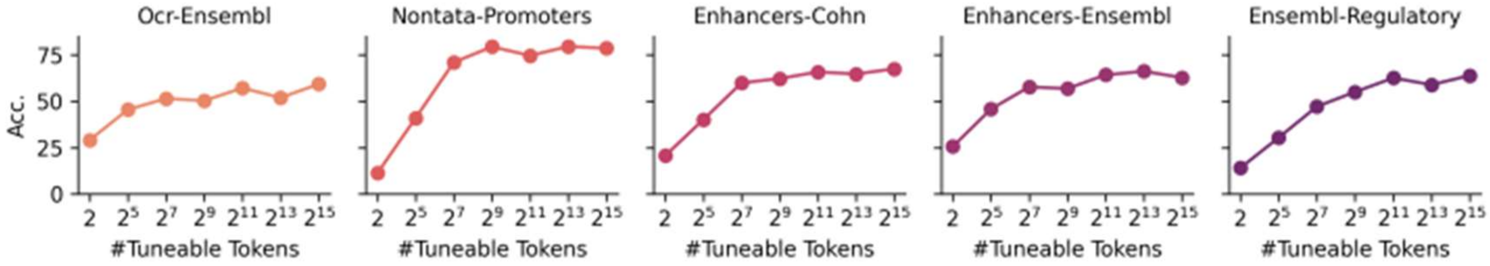


Image source: https://doi.org/10.48550/arXiv.2306.15794

In-Context Learning Capabilities:

- Few-Shot Learning Approach:
  - HyenaDNA was tested in a few-shot learning scenario, where it was given a small number of examples (k-shot) for new tasks
  - The model showed that with brief tuning, it could effectively learn and apply the concepts needed for accurate classification, even with minimal data
  - This few-shot capability is crucial in genomics, where new and unique tasks frequently arise, and having a model that can quickly adapt is invaluable

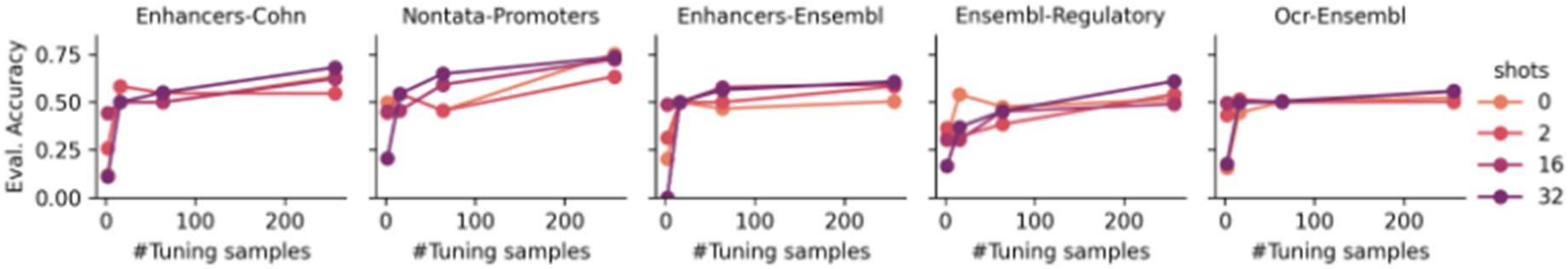
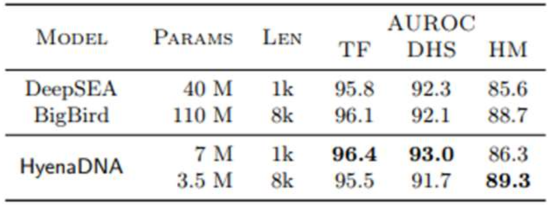


Image source: https://doi.org/10.48550/arXiv.2306.15794

Ultra-Long Range Genomics:

- Chromatin Profile Prediction:
  - HyenaDNA was applied to the DeepSEA dataset for predicting chromatin profiles
  - Despite using 5-30 times fewer parameters, HyenaDNA performed competitively with the sparse-attention BigBird model, demonstrating its ability to handle complex, multi-task genomic predictions efficiently



MODEL	PARAMS	LEN	AUROC TF	AUROC DHS	AUROC 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**

Source: https://doi.org/10.48550/arXiv.2306.15794

Ultra-Long Range Genomics:

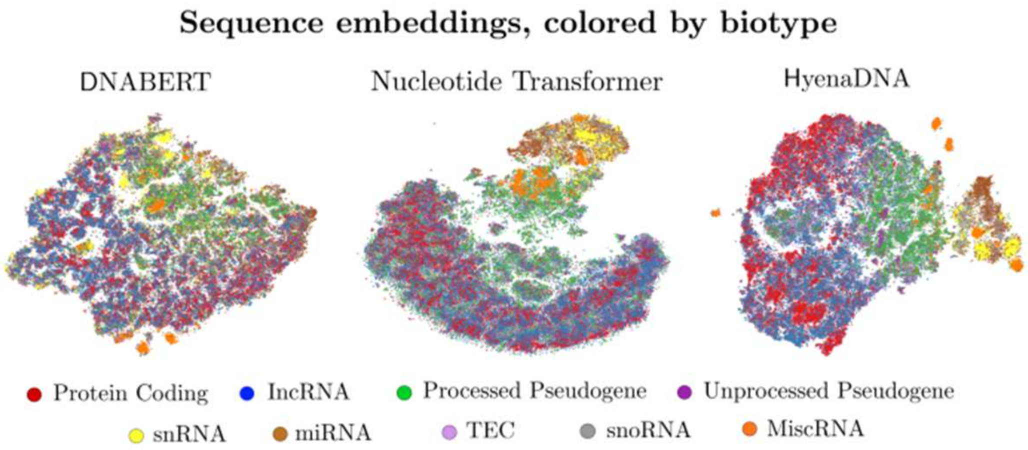
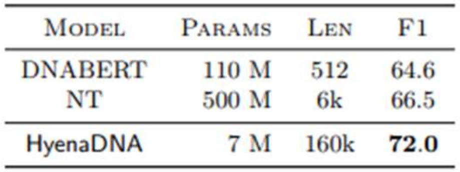


Image source: https://doi.org/10.48550/arXiv.2306.15794

- Visualization and Embedding Analysis:
  - HyenaDNA was used to generate embeddings for sequences corresponding to different biological functions
  - The embeddings produced showed clear clustering based on biotype annotations
  - HyenaDNA achieved the highest F1 score in biotype classification tasks, validating its ability as a universal genomic featurizer
  - Learn and represent features across diverse genomic tasks suggests that HyenaDNA can serve as an effective universal featurizer in genomics

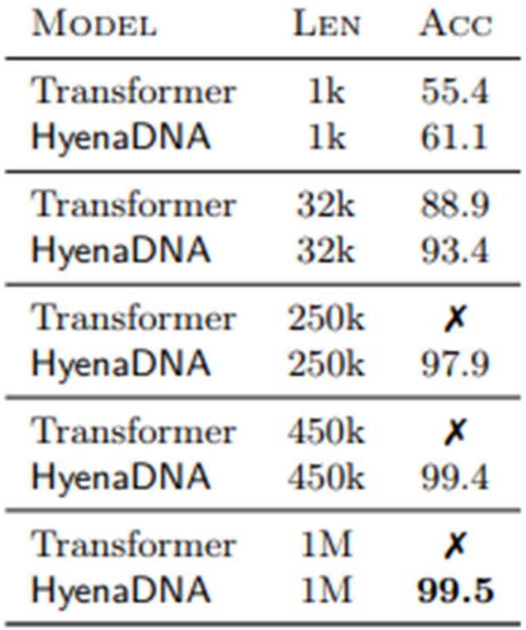


Model	Params	Len	F1
DNABERT	110 M	512	64.6
NT	500 M	6k	66.5
HyenaDNA	7 M	160k	**72.0**

Source: https://doi.org/10.48550/arXiv.2306.15794

Ultra-Long Range Genomics:

- Species Classification with Ultra-Long Sequences
  - novel species classification task, the model had to determine species origin from five different species, including humans and non-human primates
  - As the context length increased, HyenaDNA's performance improved dramatically, achieving near-perfect accuracy (99.5%) with sequences up to 1 million tokens long
  - Capacity to process ultra-long sequences and detect subtle genetic differences that distinguish species is beyond capabilities of transformer models due to infeasible training times



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**

Source: https://doi.org/10.48550/arXiv.2306.15794

Conclusion

- HyenaDNA’s innovative architecture allows it to scale efficiently with sequence length, making it capable of processing up to 1 million tokens
- Ability to adapt to tasks through in-context learning and soft prompting without retraining the entire model highlights flexibility and computational efficiency
- State-of-the-art performance on a variety of genomic tasks, outperforming larger models with far fewer parameters and less pretraining data

Future Directions

- Enhance HyenaDNA’s generalizability and reduce potential biases,
  - Pretraining the model on a more diverse set of genomes
- Extending to incorporate other biological sequences, like proteins and drug molecules,
  - Could unlock multi-modal applications
  - Allow for comprehensive modeling of the complex interactions between various biological systems
- Increasing the model size and leveraging model parallelism could push the boundaries of what HyenaDNA can achieve, allowing it to handle even longer sequences and more complex genomic tasks

Any questions?

Image Sources

- https://africafreak.com/spotted-hyena-facts
- https://www.ashg.org/discover-genetics/building-blocks/
- https://doi.org/10.48550/arXiv.2306.15794