27TH EUROPEAN CONFERENCE
ON ARTIFICIAL INTELLIGENCE
19-24 OCTOBER 2024
Santiago de Compostela

Unlocking Efficiency: Adaptive Masking for Gene Transformer Models

Soumyadeep Roy, Shamik Sural, Niloy Ganguly
Indian Institute of Technology Kharagpur

Need for annotation of non-coding DNA

- Non-coding DNA is quite large and Limited annotated datasets of non-coding DNA is available
- Gene Transformers trained in unsupervised manner on Human Reference Genome

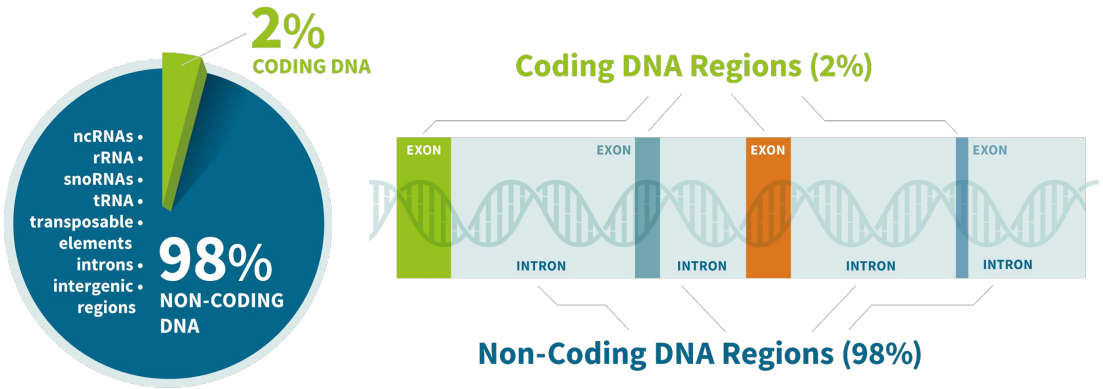
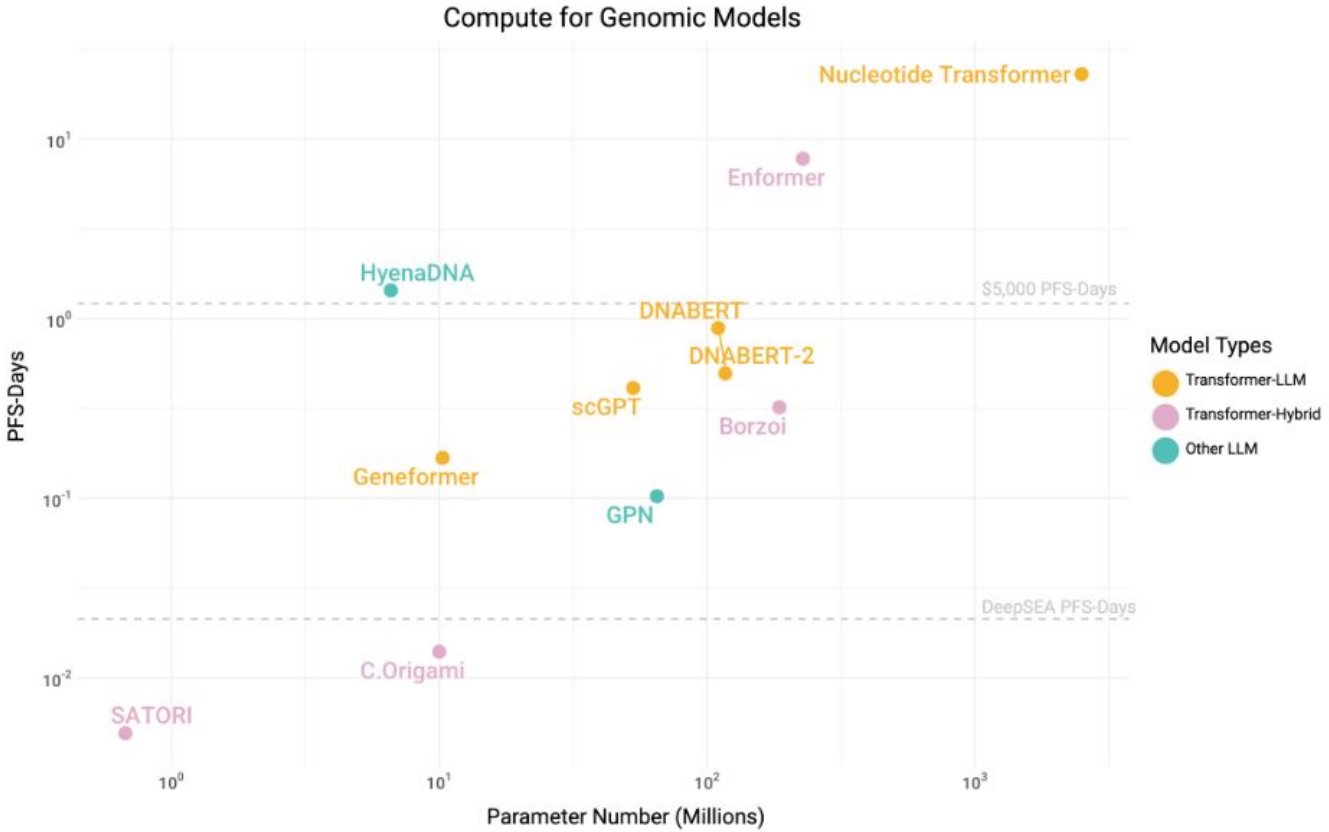


Image: https://cmsasset.ancestrycdn.com/content/dam/ancestry/magnolia-dam/seo-dna-lp-images/non-coding-dna.png



The total amount of compute, in petaflop/s-days (PFS-Days) used to train various Genomic Foundational Models (DNA, RNA, scRNA) — **Huge Computing Resources involved**

Deep Learning Models	Compute resources used for pretraining
DNABERT [Bioinformatics 2021]	25 days on 8 NVIDIA 2080Ti GPUs
DNABERT-2 [ICLR 2024]	~14 days using eight Nvidia RTX 2080Ti GPUs
GeneFormer [Nature 2023]	~3 days distributed across three nodes, each with four Nvidia V100 32GB GPUs (a total of 12 GPUs)
Enformer [Nature Methods 2021]	64 TPU v3 cores with a batch size of 64 (1 per core) for 150,000 steps (approximately 3 days)

Very costly pretraining process — **need for efficient pretraining**

Gene Transformers follow Pretraining - Finetuning

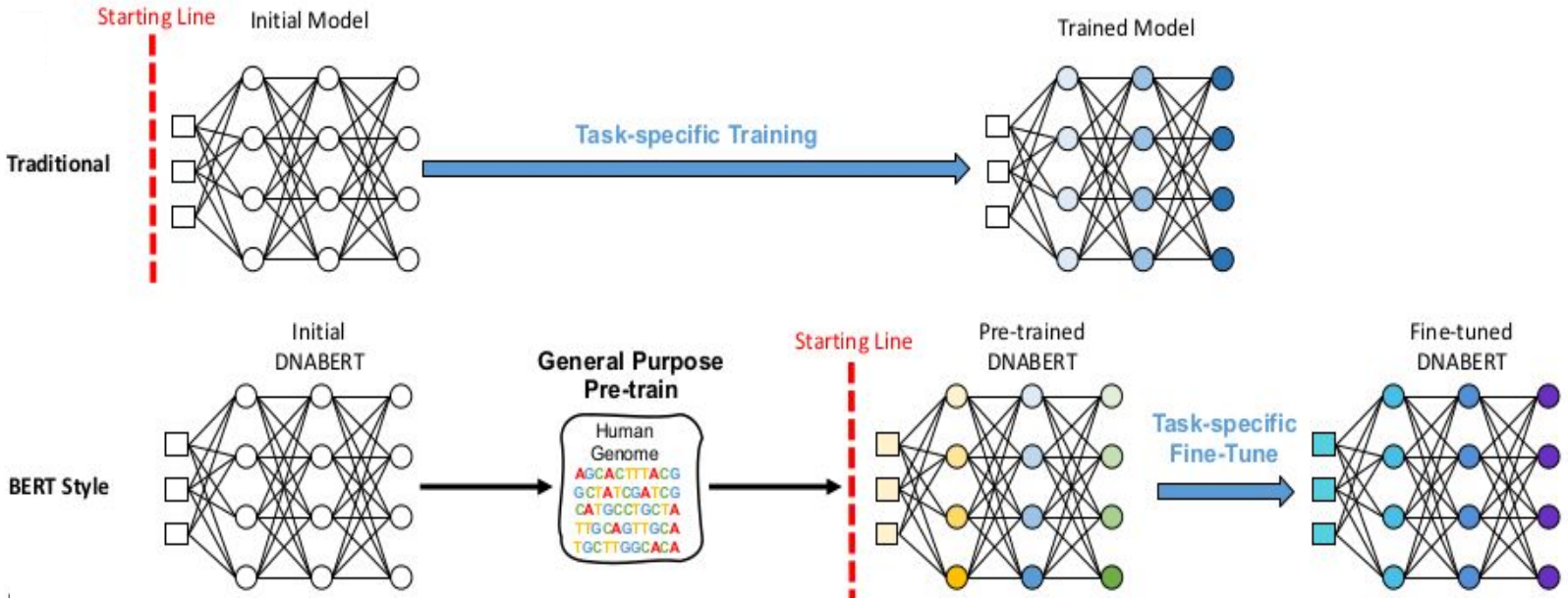
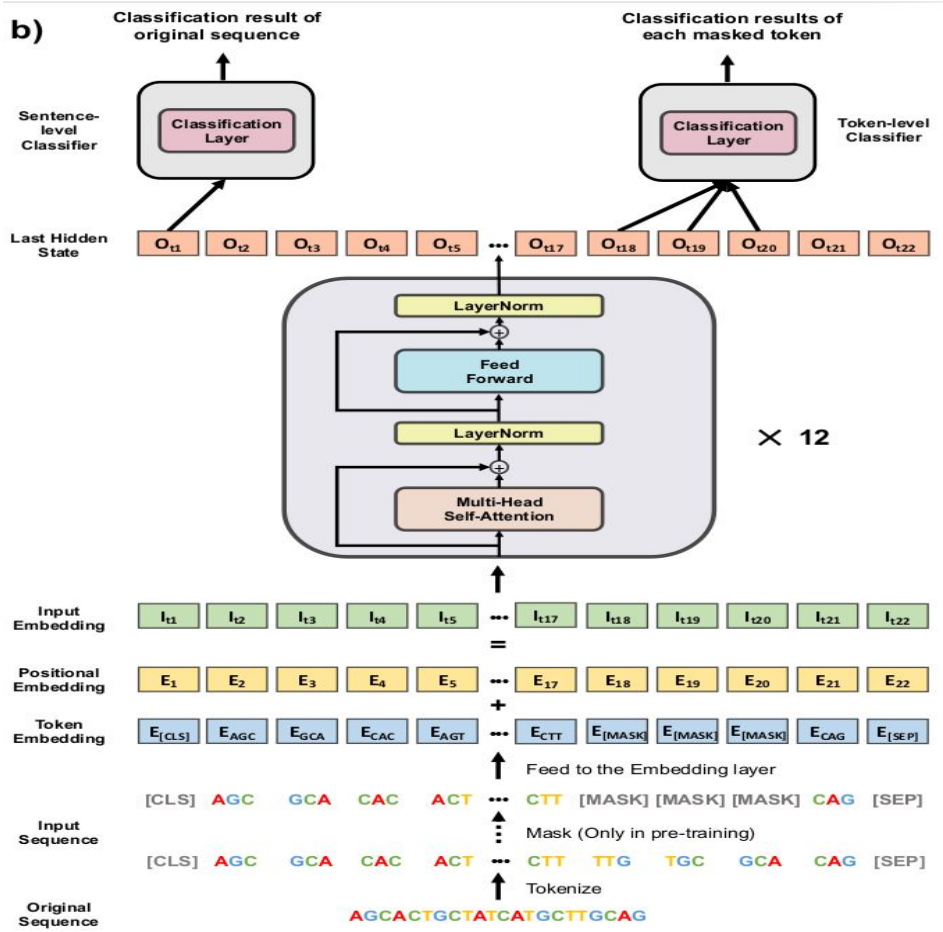


Image source: Ji et al. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, Bioinformatics, pp. 1-9

Base model - DNABert

- DNABert - Combines BERT with DNA sequences of 6-mers (ATTCGC)
  - Model vocabulary size for **6-mer model: 4^6 plus special tokens**

- Gene regulatory code (non-coding) is complex, shows signs of **polysemy, distant semantic relationship** between sequence codes
  - Cis-regulatory elements acts similar to language


b)
Classification result of original sequence
Classification results of each masked token
Sentence-level Classifier
Classification Layer
Classification Layer
Token-level Classifier
Last Hidden State
LayerNorm
Feed Forward
LayerNorm
Multi-Head Self-Attention
× 12
Input Embedding
Positional Embedding
Token Embedding
Feed to the Embedding layer
[CLS] AGC GCA CAC ACT ••• CTT [MASK] [MASK] [MASK] CAG [SEP]
Input Sequence
Mask (Only in pre-training)
[CLS] AGC GCA CAC ACT ••• CTT TTG TGC GCA CAG [SEP]
Tokenize
Original Sequence
AGCACTGCTATCATGCTTGCAG

Gene Transformers - DNABert

- Standard BERT architecture with 12 attention blocks

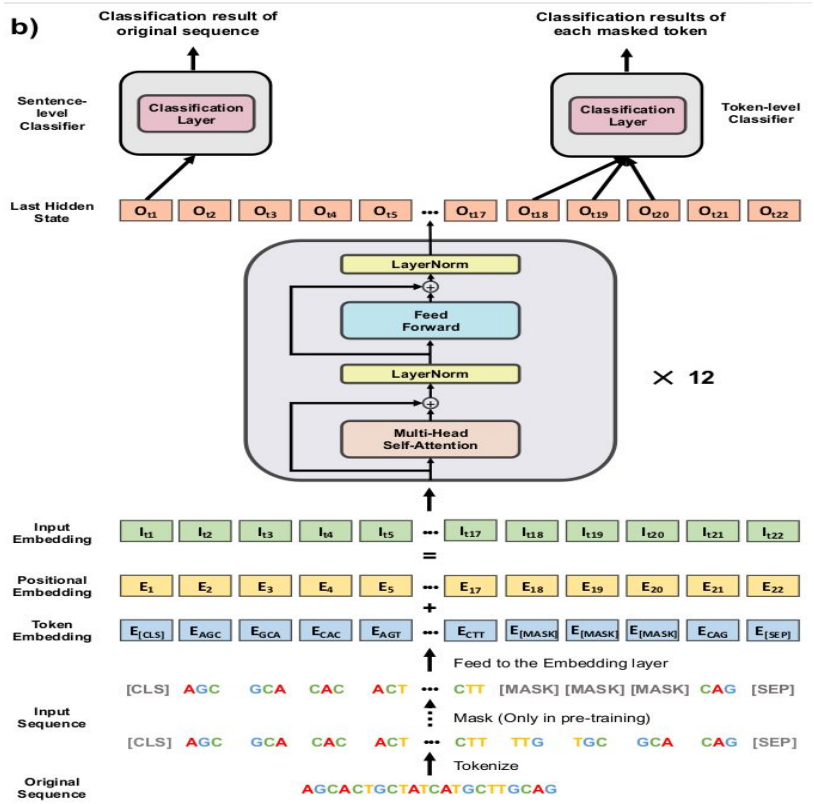


Image source: Ji et al. (2021) DNABERT

Gene Transformers - DNABert

- Standard BERT architecture with 12 attention blocks
- No word or sentence-level information exist for gene sequences
- Tokens: 6-mers (like AGCGCA)
  - Example of 3-mers is given in the figure

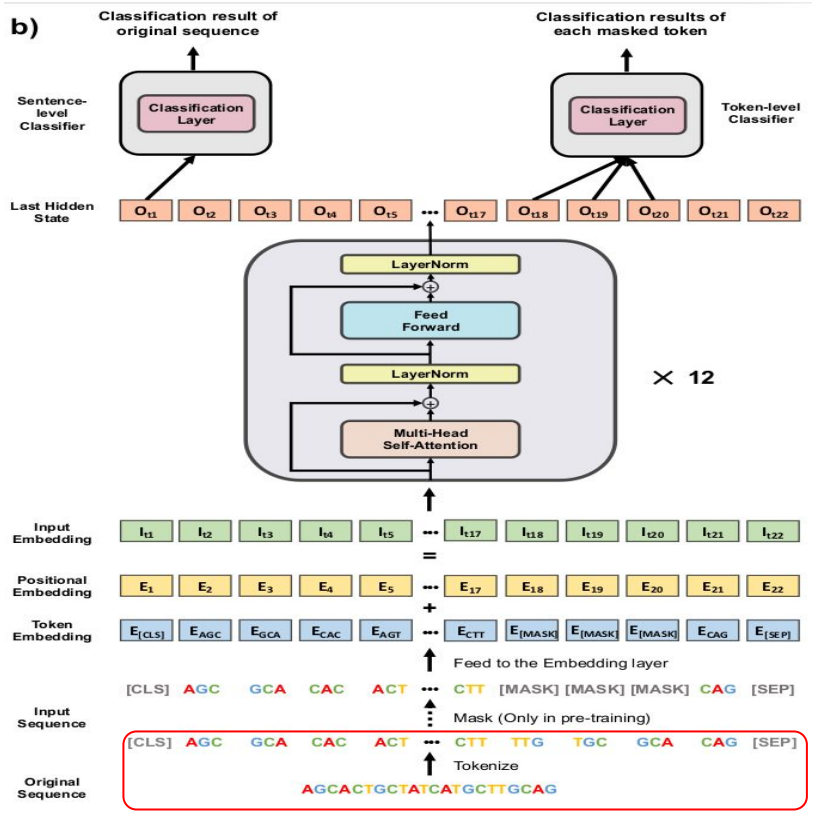


Image source: Ji et al. (2021) DNABERT

Problem Statement

To achieve comparable or better accuracy on gene sequence classification tasks by efficient pretraining of genomic foundational models

SOTA model achieves an accuracy of **X** by pretraining on **N** steps

Aim: Improve MLM Masking achieves accuracy of Y by pretraining on M steps

$$M << N$$

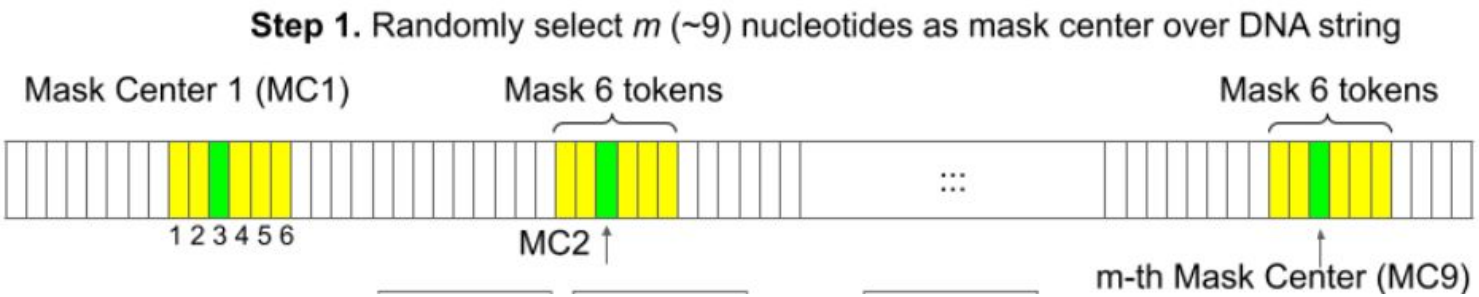
$$Y >= X$$

Replace Random Span Masking with PMI Masking

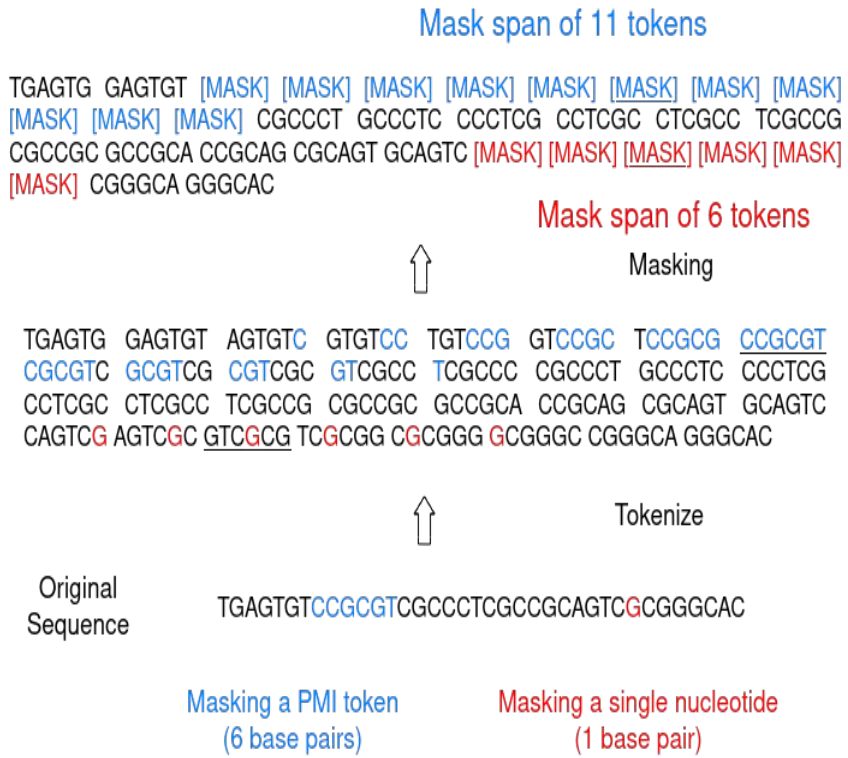
- Random masking allows abusing local features (not learning the overall context)
    - the United [MASK] : the United States
    - by [MASK] way : by the way
  - Training steps are wasted for "easy" predictions
- Words in NLP = _____ in gene sequences
  - Difficult to identify semantic-preserved tokens

- **Idea: Jointly mask multiple tokens if they exhibit high collocation**

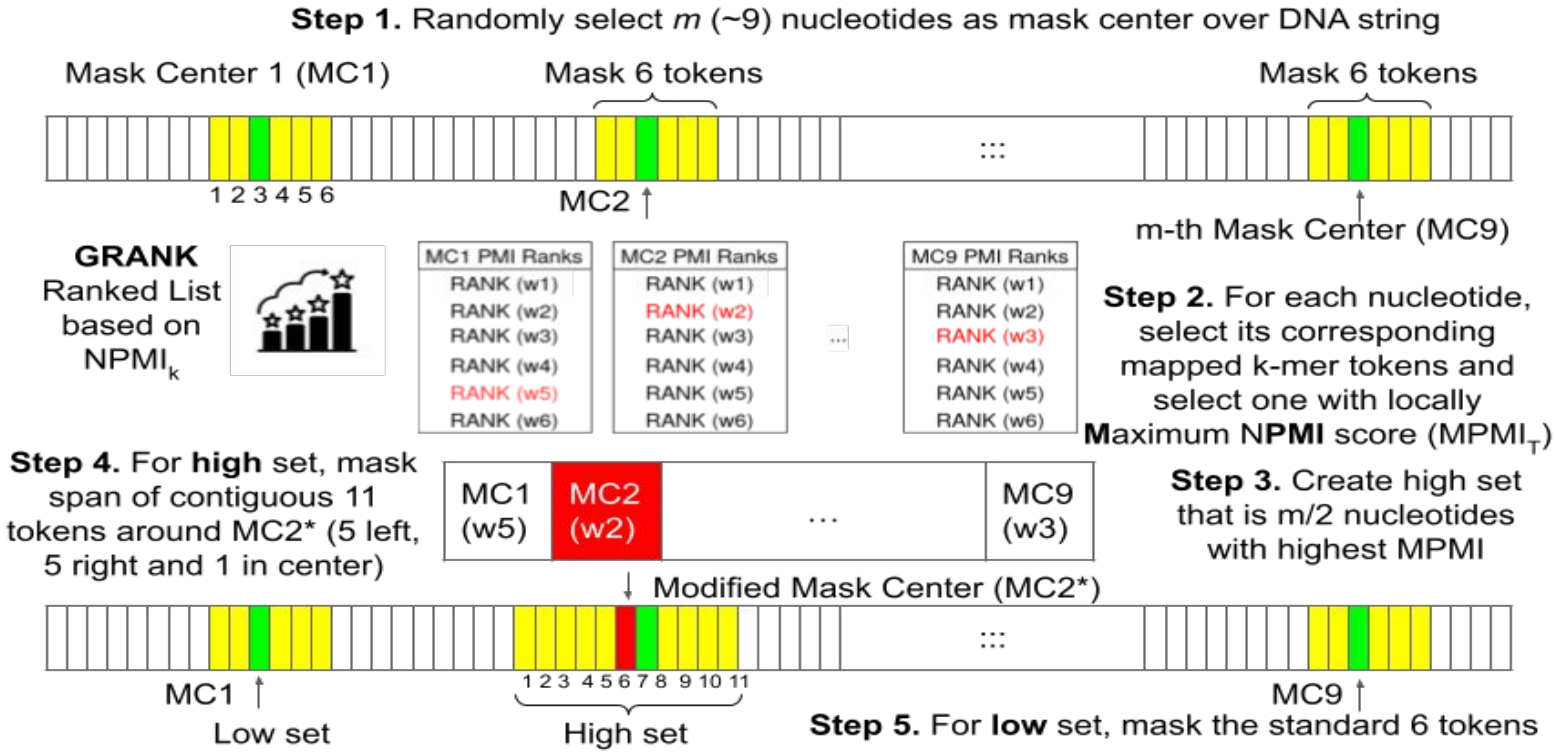
Random Span Masking in Action



Masking a nucleotide means masking six tokens



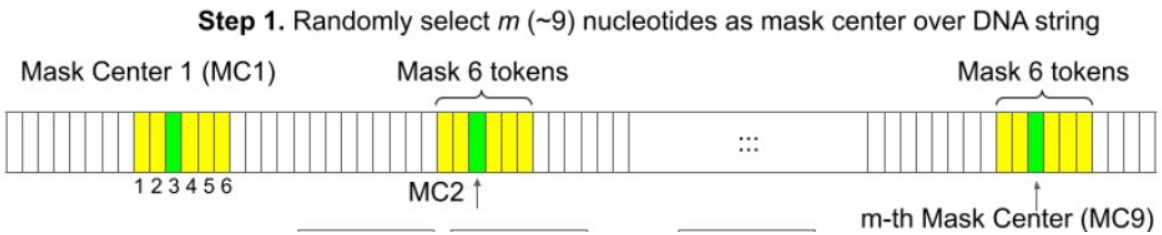
Research Background: GeneMask Algorithm



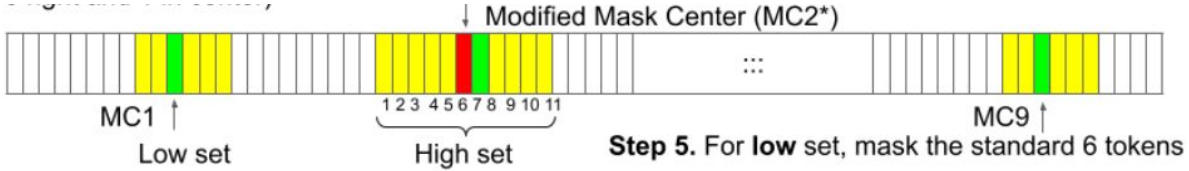
GENEMASK: Fast Pretraining of Gene Sequences to Enable Few-Shot Learning, ECAI 2023

Proposed Global PMI Masking

Random Span Masking (DNABert)



GeneMask (our ECAI 2023 paper)



Global-PMI Model (proposed)

Only mask the (globally) high PMI tokens, without taking from low set

Proposed Methodology - Global PMI Masking

Algorithm 1: GLOBAL Algorithm

Input: Input sequence of 6-mer tokens having a maximum of 510 tokens, Pre-computed Normalized PMI_k ($NPMI_k$) values for all 6-mers stored as a dictionary

Output: $MaskTokenSet$: Token indices within input sequence to be masked

Initialization: `// A 6-mer token present at` i`-th position is represented as T[i], the` i`-th nucleotide is represented as DNA[i]`

$MaskTokenSet \leftarrow \emptyset$

T[i] $\leftarrow$ {DNA[i-2] $\cdots$ DNA[i+3]}

Function `MapNucleotideToKmerTokens` (*nucleotide position id* i) :

  $MappedTokens \leftarrow$ T[j], $\forall(j)_{j=i-2}^{i+3}$

  return $MappedTokens$

Step 1: Sort the DNA string with 6-mer tokens in a non-increasing order of $NPMI_k$ score.

Step 2: Create a priority set with the top-ranked m nucleotides with the highest $NPMI_k$ score.

Step 3: **for** *each nucleotide in priority set* **do**

  `// Masking a PMI token involves masking 11 adjacent tokens`

  $MaskTokenSet \leftarrow MaskTokenSet \cup \forall(j)_{j=\tau-2}^{\tau+3}$ $MapNucleotideToKmerTokens$(j)

return $MaskTokenSet$

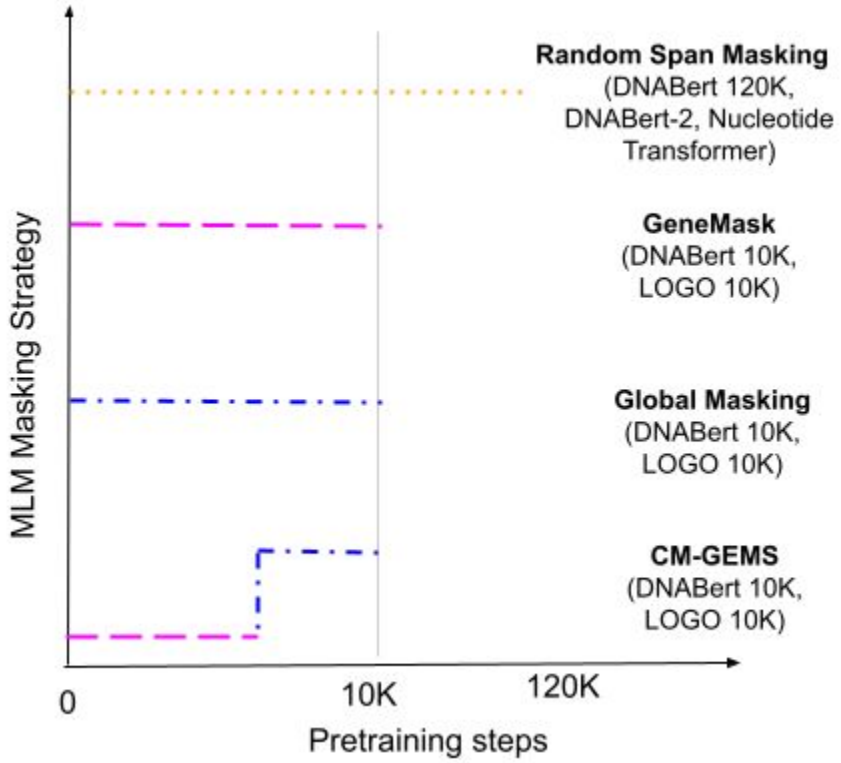
Improve Global PMI Masking by making it dynamic

- Increase the percentage of high PMI tokens among the masked tokens
  - Propose “Global-PMI” masking method

Proposed Curriculum Masking (CM-GEMS)

- Masking strategy changes during the pre-training steps
- Easy strategy at the start (GeneMask strategy). When drop in perplexity score drops below one, the masking strategy changes to “Global” strategy (Hard strategy)
- Easy strategy : GeneMask, Hard Strategy : Proposed Global PMI Masking

Comparison of MLM Masking Strategies



Datasets: Few-shot Setting

- Promoter Region Prediction - binary classification (two tasks)
  - Prom-core: -35 bp to +34 bp around TSS
  - Prom-300: -249 bp to +50 bp around TSS

- Enhancer prediction - 500 bp
  - An enhancer is a sequence of DNA that can bound specific proteins and therefore increase a change of transcription of a particular gene. Unlike promoters, enhancers do not need to be in a close proximity to TSS (might be several Mb away

Datasets: Few-shot Setting

- Splice Donor and Acceptor Site Prediction - predict whether donor, acceptor or non-splice site (3-way classification) - 40 bp
  - Extract 40 bp long sequence around the donor and acceptor sites of exons as positive sequences

- Silencer Prediction - 300 bp
  - A silencer is a DNA sequence capable of binding transcription regulation factors, called repressors
  - Silencers prevent genes from being expressed as proteins

Datasets: GUE Benchmark (ICLR 2024) Full Data

- Consists of **seven** gene sequence classification tasks with **28 datasets** with input sequence lengths ranging between 70 to 1000
    - Did not evaluate on Covid dataset

- Evaluating multi-species generalizability: 15 out of the 27 datasets of GUE belong to Non-human species
    - 5 datasets from Mouse for the Transcription Factor Prediction task
    - 10 datasets from Yeast for the Epigenetic Marks Prediction task

DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome (ICLR 2024)

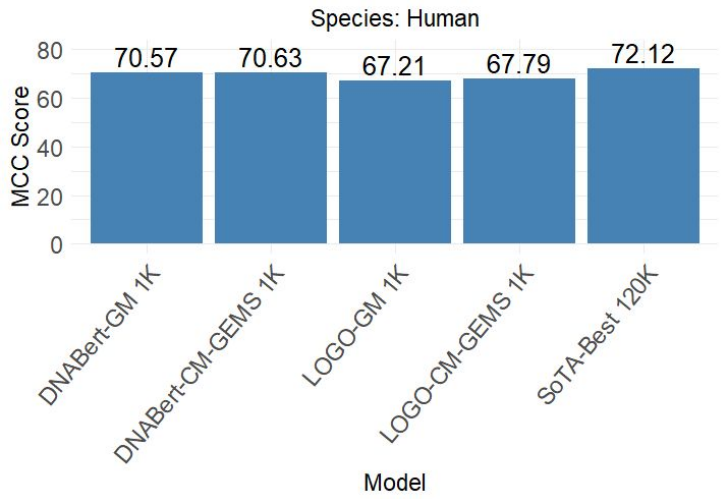
Experimental Setup

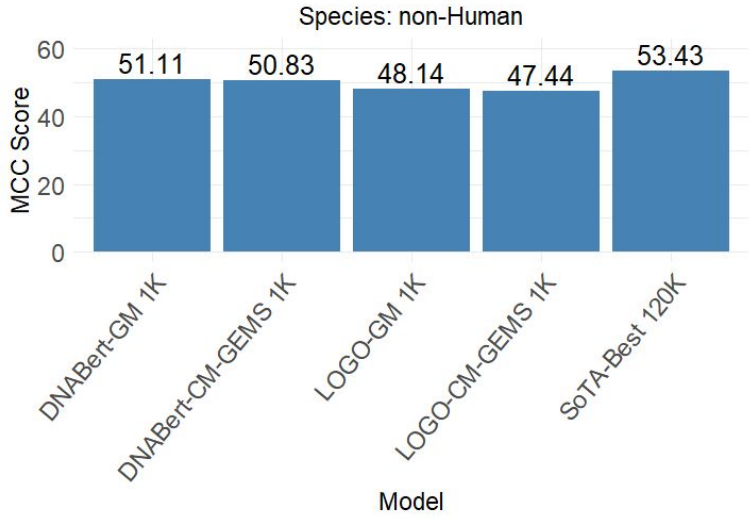
- Evaluation Metric
  - Accuracy for few-shot setting - Few-shot setting by Roy et al. [ECAI 2023]
  - Matthew's correlation coefficient - GUE Benchmark [ICLR 2024]
- Baseline Models
  - Recent Foundational Models - **DNABERT, DNABERT-2, Nucleotide Transformer**
  - Curriculum Masking baseline - Divide total training steps by ten and increase percentage of Global Masking algorithm and reduce the GeneMask masking algorithm

CM-GEMS 10K performs best on Genome Understanding Evaluation Benchmark

Task	Dataset	Models											
		State-of-the-art (SoTA) models				DNABert				LOGO			
		DNABert-2 120K (BPE)★	DNABert 120K◇	NT-500M-Human◇	DNABert-2 120K (k-mer)★	Global 10K	GeneMask 10K	CM-Step 10K	CM-GEMS 10K	Global 10K	GeneMask 10K	CM-Step 10K	CM-GEMS 10K
Species: Human													
PD (Human)	all	85.57	90.48	87.71	83.78	89.93	89.50	90.48	89.29	85.82	82.93	85.88	84.76
	no tata	92.55	93.05	90.75	92.65	91.09	91.73	91.83	92.41	89.52	88.54	89.79	88.62
	tata	60.85	61.56	78.07	57.75	76.25	79.76	81.12	77.47	68.24	69.79	69.31	72.21
CPD (Human)	all	66.28	68.90	63.45	74.91	70.56	68.54	71.65	72.09	69.16	64.13	67.34	63.66
	notata	67.99	70.47	64.82	69.23	70.87	70.01	72.13	70.18	68.91	67.29	66.84	66.23
	tata	72.73	76.06	71.34	74.91	76.96	74.95	75.51	83.50	70.85	53.65	55.55	61.71
TFP (Human)	0	66.99	66.84	61.59	67.99	65.89	67.40	65.44	66.07	64.21	67.09	63.94	65.34
	1	70.98	70.14	66.75	67.06	71.14	69.81	68.35	68.95	69.47	67.85	67.49	69.75
	2	61.40	61.03	53.58	59.45	57.66	59.80	58.61	57.66	53.37	55.31	53.83	55.31
	3	55.10	51.89	42.95	50.24	46.80	51.26	47.65	51.41	40.20	42.48	44.70	40.49
	4	71.31	70.97	60.81	72.80	74.10	76.15	73.22	72.60	70.32	70.54	68.65	69.98
Splice	Reconstruct	79.62	84.07	79.71	77.90	83.02	84.84	84.74	84.12	77.92	74.01	80.25	75.20
Mean (Human)		70.95	**72.12**	68.46	70.72	72.86	73.65	73.39	**73.81**	**69.00**	66.97	67.80	67.77
Species: non-Human													
EMP (Yeast)	H3	77.08	73.10	69.67	74.62	71.45	73.28	74.07	74.35	64.72	60.90	61.91	61.49
	H3K14ac	55.60	40.06	33.55	42.71	38.75	40.73	40.27	41.28	30.38	32.55	33.34	29.49
	H3K36me3	57.25	47.25	44.14	47.26	44.11	45.42	46.06	46.83	38.94	39.26	38.52	35.92
	H3K4me1	45.51	41.44	37.15	39.66	41.63	42.36	40.91	44.61	31.19	28.66	31.04	25.38
	H3K4me2	40.83	32.27	30.87	25.33	30.60	33.14	31.40	33.43	30.99	29.32	30.11	27.11
	H3K4me3	42.57	27.81	24.00	27.43	25.91	25.92	24.93	30.24	18.34	15.35	22.65	12.22
	H3K79me3	66.01	61.17	58.35	61.03	59.18	60.45	59.20	59.83	54.36	52.19	55.70	53.38
	H3K9ac	56.79	51.22	45.81	49.35	49.24	52.22	51.77	52.77	43.54	42.16	45.62	40.04
	H4	80.07	79.26	76.17	78.61	76.38	76.04	75.83	76.66	72.81	66.51	71.51	68.85
	H4ac	54.19	37.43	33.74	37.14	33.89	37.43	35.69	36.21	27.76	27.84	31.84	27.5
TFP (Mouse)	0	48.01	44.42	31.04	48.96	49.48	52.57	50.48	54.32	12.02	27.93	27.16	42.23
	1	81.86	78.94	75.04	81.69	79.70	79.05	79.90	80.51	71.30	69.73	70.93	69.50
	2	82.98	71.44	61.67	81.71	75.50	78.08	74.40	80.70	52.50	59.80	55.57	78.19
	3	73.22	44.89	29.17	63.17	51.00	60.27	52.51	61.01	34.87	48.96	34.38	61.53
	4	46.15	42.48	29.27	42.83	41.04	42.60	42.68	44.94	21.40	28.78	25.89	26.60
Mean (non-Human)		60.54	51.55	45.31	**53.43**	51.19	53.30	52.01	**54.51**	40.34	42.00	42.41	**43.96**

CM-GEMS at 1K steps versus SoTA at 120K



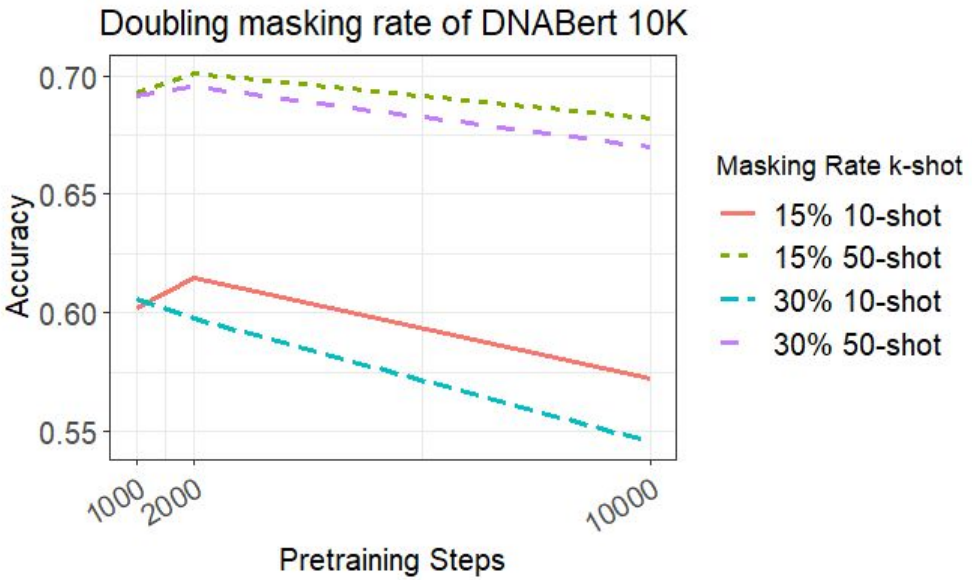


CM-GEMS 1K outperforms GeneMask (GM) 1K for Human species only

CM-GEMS achieves 90% of the SoTA-best 120K model, except for LOGO 1K for non-Human species

Effect of doubling the masking rate

- Performance **drop** at 10K steps over the standard masking rate of 15% is4.71% and 1.79% respectively
- Higher masking rate, the model sees a **much-reduced** amount of unmasked or actual tokens that hinder the learning process.



Conclusion

- Limitations of conventional tokenization methods in gene transformers models
- Proposes a novel curriculum masking approach to address these shortcoming by systematically increasing hardness of masked token prediction task
- Evaluated on 32 tasks and with SOTA models such as DNABert, DNABert-2 (ICLR 2024), Nucleotide Transformer
- Highly efficient pretraining strategy that is generalizable to settings without known grammar, i.e., non-language strings

Future Work

- Extend our work to **other foundational models based on genomics data**, specifically RNA-based models
    - CodonBERT, RNABERT and single-cell RNA-based models such as GeneFormer

- Although PMI indirectly captures DNA sequence motifs, a much-needed inter-disciplinary research direction is to **involve more biologically grounded pretraining or fine-tuning objectives** instead of **MLM**

Acknowledgments

- Institute Ph.D. Fellowship at the Indian Institute of Technology Kharagpur
- Conference travel expenses is partially provided from Project UTT of IIT Kharagpur and IARCS-ACM India
- Complex Networks Research Group, IIT Kharagpur, India

@cnerg

Thank you for your attention

GitHub: https://github.com/roysoumya/curriculum-GeneMask

Paper: https://ebooks.iospress.nl/doi/10.3233/FAIA240864

Contact me at:  soumyadeep.roy9@iitkgp.ac.in

 @roysoumya1