DNA-MSAT: DNA-based MSA Transformer Advances DNA Feature Prediction

Sukhwan Park[1] and Martin Steinegger[1,2,3,4,*]

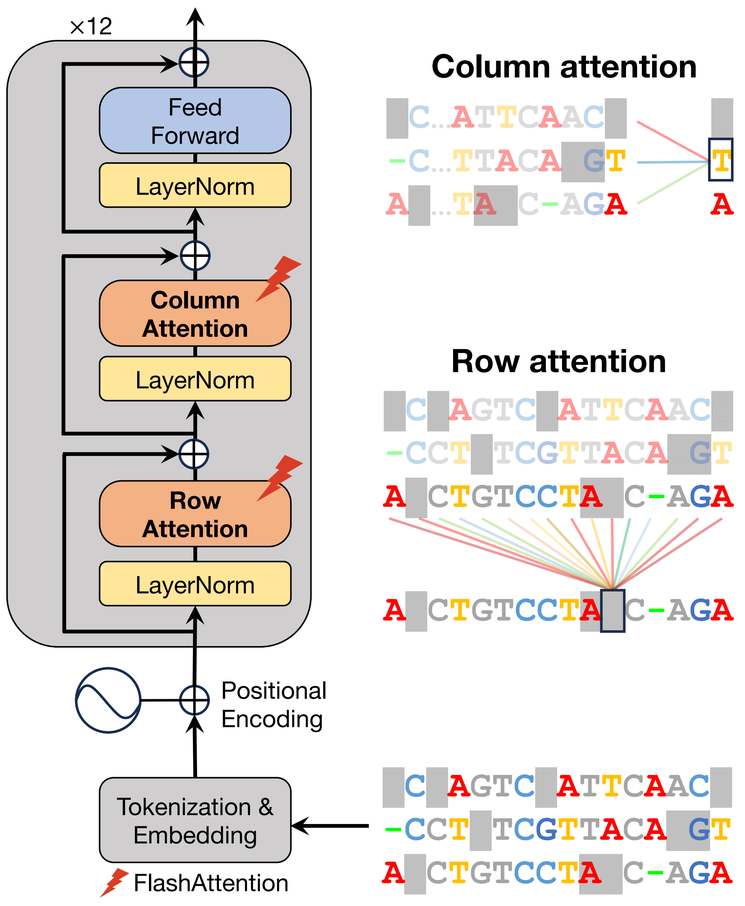
[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Republic of Korea, [2]School of Biological Sciences, SNU, [3]Institute of Molecular Biology, SNU, [4]Artificial Intelligence Institute, SNU. [*]Correspondence: martin.steinegger@snu.ac.kr

Abstract

Transformer-based language models deliver state-of-the-art performance in diverse bioinformatics tasks ranging from protein structure prediction to DNA feature identification. A significant gap persists in current DNA language models: the underutilization of homology signals derived from whole-genome Multiple Sequence Alignments (MSAs). To exploit the rich evolutionary information contained in these, we trained a novel **DNA-based MSA Transformer (DNA-MSAT)** on clustered whole-genome MSAs of 100 vertebrate species with the human genome as reference. We show that DNA-MSAT embeddings have higher discriminative power on DNA features (e.g. transcription factor binding sites, silencers, promoters, ...) relative to other embeddings (**Table 1**). After fine-tuning for **coding region prediction**, DNA-MSAT accuracy of **97.6%** outperformed the 20x larger Nucleotide Transformer's accuracy of 95.2%.

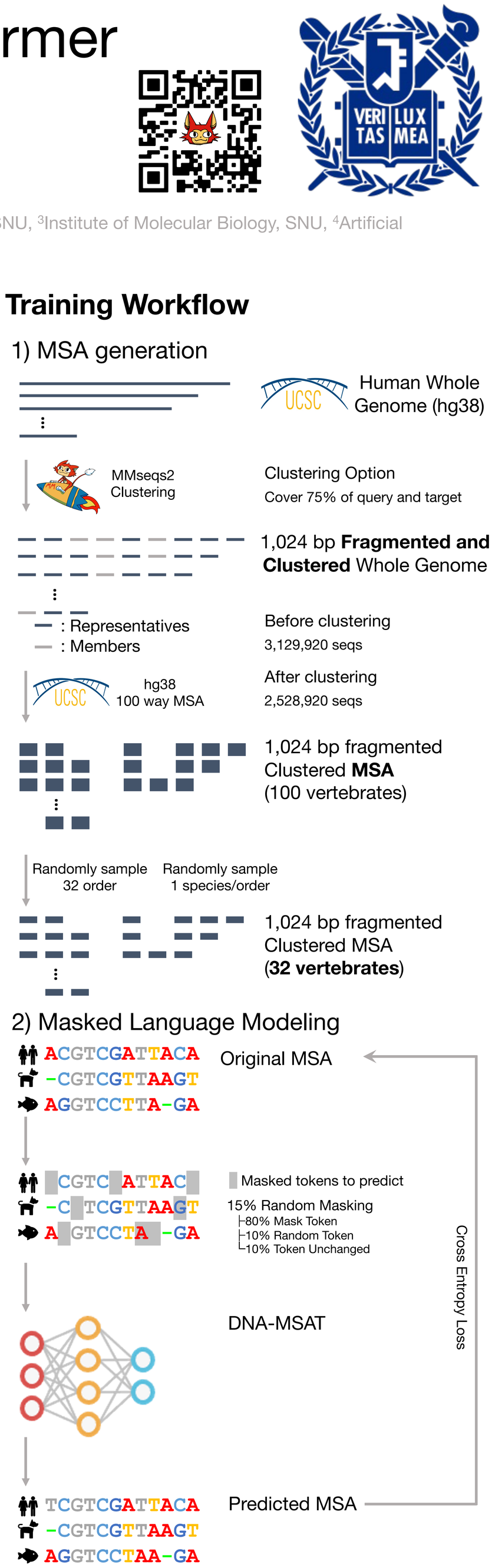
We are extending DNA-MSAT to other DNA annotation tasks and plan to deliver a comprehensive free and open-source platform to improve our understanding of both well-studied model organisms and the whole tree-of-life.

MSA Transformer Architecture



Row and Column attention efficiently calculate the attention over the MSA and reduce memory complexity from $O(M^2L^2)$ to $\boldsymbol{O(M^2L) + O(ML^2)}$. Row attention helps MSA information utilization. FlashAttention speeds up training and inference.

Training Workflow



Benchmark

	DNA-MSAT (Ours)	GPN-MSA	Nucleotide Transformer	DNABERT2
# Parameters	115M	86M	2.5B	117M
Coding Region				
1,024bp	**0.952**	0.937	0.882	0.844
128bp	**0.974**	0.965	0.842	0.826
Transcription Factor Binding Site: Difficult				
POLR2A, 1,020bp	0.441	0.316	**0.466**	0.455
POLR2A, 128bp	0.464	0.418	0.423	**0.466**
GATA3, 1,020bp	0.333	0.111	0.439	**0.478**
GATA3, 128bp	0.294	0.350	**0.401**	0.400
MAFK, 1,020bp	**0.656**	0.019	0.514	0.616
MAFK, 128bp	0.393	**0.539**	0.432	0.444
Transcription Factor Binding Site: Easy				
POL2, 1,020bp	0.770	0.764	**0.800**	0.789
POL2, 128bp	0.890	0.881	**0.899**	0.893
BPOL2, 1,020bp	0.794	0.802	**0.856**	0.855
BPOL2, 128bp	**0.871**	0.812	0.857	0.851
TAF1, 1,020bp	0.817	0.829	**0.848**	0.829
TAF1, 128bp	0.813	**0.842**	0.831	0.830

Table 1: Matthews Correlation Coefficient (MCC) for each prediction task.
Bold shows the best performing model for each task.
10 epochs, batch size of 32, 16,000 training, 1,000 validation, and 1,000 test samples
Checkpoints saved every 200th step, evaluated with the lowest validation loss on the test dataset

References

Rao, et al. PMLR, 139, 8844-8856, 2021.
Lee, et al. *NAR* 48.D1, 2020.
Steinegger & Söding. *Nature Biotechnology* 35.11, 2017.
Dao, et al. *NeurIPS* 35, 2022.
Zhou, et al. *arXiv preprint arXiv:2306.15006,* 2023.
Dalla-Torre, et al. *bioRxiv,* 2023.01.11.523679, 2023.
de Souza. *Nature Methods* 9.11, 2012.

Acknowledgments

Artificial Intelligence Institute
Seoul National University