DNABERT-2: EFFICIENT AND EFFECTIVE FOUNDATION MODEL FOR MULTI-SPECIES GENOME

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta,
Ramana V. Davuluri, Han Liu

DNABERT

Understand genome is a fundamental task in biology research.

Various important sub-task:

- Promoter Detection
- Splice Site Prediction
- Transcription Factor Prediction
- Epigenetic Marks Prediction
- ...

DNABERT

Pre-train of DNABERT

Human Genome

ATGTCAATGTCAATGTCAATGTCAACTGTCAATTACTGTCAATTACTGTCAATTGCACTGTCAGACTGTCAACTGTCAATT
CAAATGTCAATTGCACTGTCAGACTGTCAATTACTGTCAATTACTGTCAATTGCACTGTCAGACTGTCAACTGTCAATTTT
GTCAATTACTGTCAATTACTGTCAATTGCACTGTCAGACTAATGTCAATTGCACTGTCAGACTGTCAATTACTGTCAGACTG

DNA sequence

ATTGCACTGTCAG

Tokenization

k-mer sequence

ATTGCA TTGCAC TGCACT GCACTG CACTGT ACTGTC CTGTCA TGTCAG

Masking

masked k-mer sequence

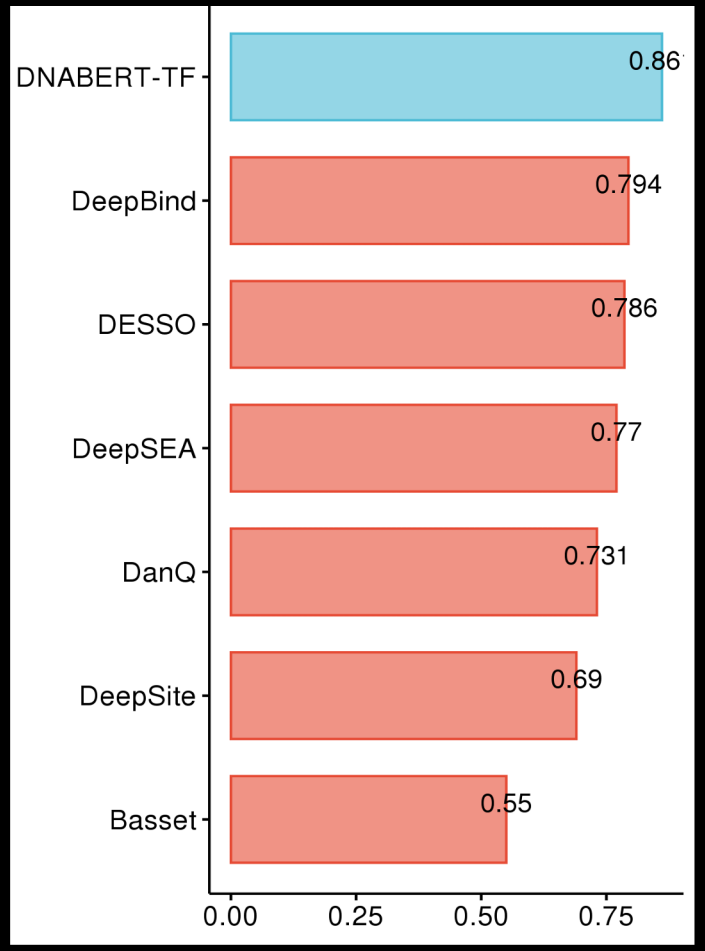
ATTGCA TTGCAC [MASK] [MASK] [MASK] [MASK] CTGTCA TGTCAG

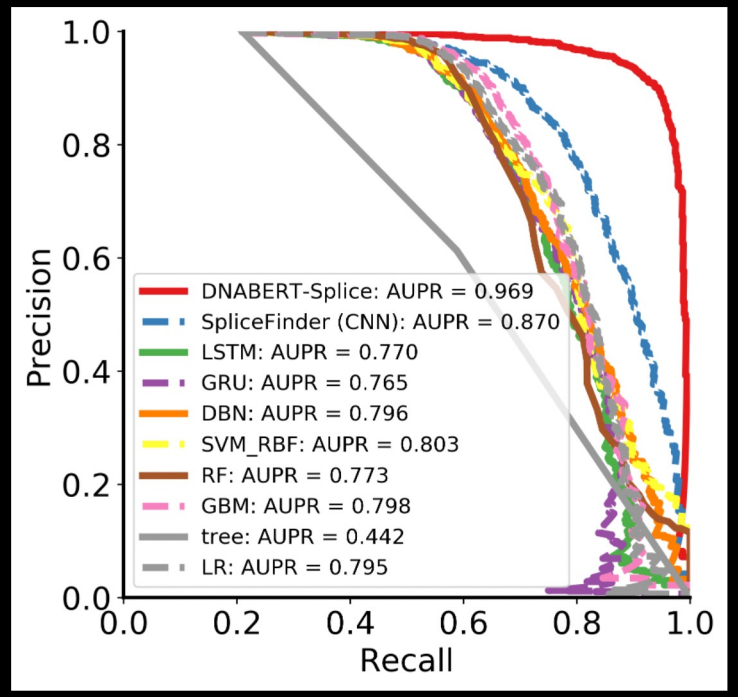
DNABERT

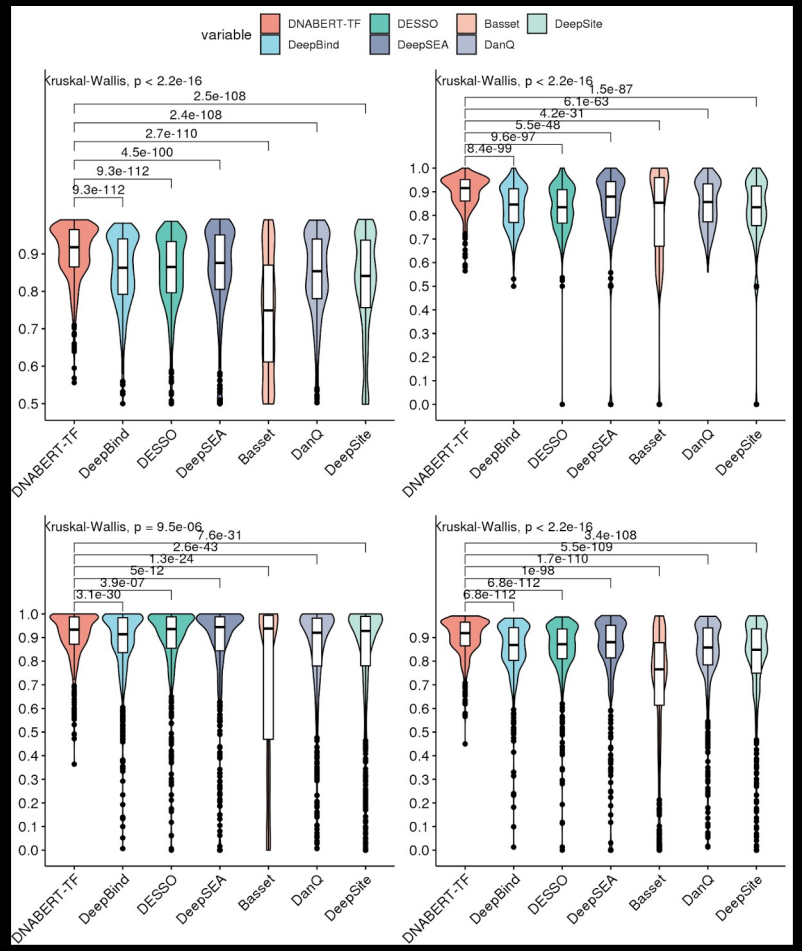
Predict Masked k-mers

TGCACT GCACTG CACTGT ACTGTC

DNABERT


DNABERT-TF
DeepBind
DESSO
DeepSEA
DanQ
DeepSite
Basset
0.86
0.794
0.786
0.77
0.731
0.69
0.55
0.00
0.25
0.50
0.75


Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
DNABERT-Splice: AUPR = 0.969
SpliceFinder (CNN): AUPR = 0.870
LSTM: AUPR = 0.770
GRU: AUPR = 0.765
DBN: AUPR = 0.796
SVM_RBF: AUPR = 0.803
RF: AUPR = 0.773
GBM: AUPR = 0.798
tree: AUPR = 0.442
LR: AUPR = 0.795


variable
DNABERT-TF
DeepBind
DESSO
DeepSEA
Basset
DanQ
DeepSite
Kruskal-Wallis, p < 2.2e-16
Kruskal-Wallis, p = 9.5e-06
DNABERT-TF
DeepBind
DESSO
DeepSEA
Basset
DanQ
DeepSite

DNABERT-2

We identify 3 key limitations of DNABERT:

1. The k-mer tokenization is data and computational inefficient
2. The model is trained on human genome only
3. The model architecture is less optimal
    - Input length limitation
    - Weak representation capability
    - …

DNABERT-2

Limitation 1: Inefficiency of k-mer tokenization

Sequence	*Token 1*	*Token 2*	*Token 3*
ATTGCACT →	ATTGCA	TTGCAC	TGCACT
	ATTGCA	[MASK]	TGCACT

[MASK]: starts with TGCAC and ends with TTGCA

Data Inefficiency

Sequence								
ATTGCACTGTCAG →	ATTGCA	TTGCAC	TGCACT	GCACTG	CACTGT	ACTGTC	CTGTCA	TGTCAG
	ATTGCA	[MASK]	[MASK]	[MASK]	[MASK]	[MASK]	[MASK]	TGTCAG

starts with TTGCA. Search space 4096 -> 4

Solution

Use BPE to replace k-mer tokenization.

Sequence	*Token 1*	*Token 2*	*Token 3*
ATTGCACTGTCAG →	ATTG	CACTG	TCAG

High-frequently co-occur segments

DNABERT-2

Limitation 1: Inefficiency of k-mer tokenization

Computational Inefficiency

With k-mer tokenization, a DNA sequence with length N will end up with (N-5) tokens.

It's computational heavy considering the O(n^2) computational efficiency of Transformer

Solution

Use BPE to replace k-mer tokenization.

- Reduce sequence length by 5 times on average
- No information leakage in pre-training

DNABERT-2

Limitation 2: Only trained on human genome

Solution Train the model on the combination of genome from 135 different species.

Limitation 3: Less optimal architecture

Solution

- Replace positional embedding with ALiBi. Now support unlimited input length.
- Incorporate Flash Attention to improve efficiency.
- Use new activation function and other tricks to improve effectiveness.

DNABERT-2

Genome Evaluation Benchmark

Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Core Promoter Detection	3	2	70
	Transcription Factor Prediction	5	2	100
	Promoter Detection	3	2	300
	Splice Site Detection	1	3	400
Mouse	Transcription Factor Prediction	5	2	100
Yeast	Epigenetic Marks Prediction	10	2	500
Virus	Covid Variant Classification	1	9	1000

Table 1: Summarization of the Genome Understanding Evaluation (GUE) benchmark.

Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Enhancer Promoter Interaction	6	2	5000
Fungi	Species Classification	1	25	5000
Virus	Species Classification	1	20	10000

Table 2: Summarization of the Genome Understanding Evaluation Plus (GUE^{+}) benchmark.

DNABERT-2

Results on GUE

Size Efficiency Effectiveness

Model	Num. Params. ↓	FLOPs ↓	Trn. Tokens	Num. Top-2 ↑	Ave. Scores ↑
DNABERT (3-mer)	86M	3.27	122B	2 || 0	61.62
DNABERT (4-mer)	86M	3.26	122B	0 || 1	61.14
DNABERT (5-mer)	87M	3.26	122B	0 || 1	60.05
DNABERT (6-mer)	89M	3.25	122B	0 || 1	60.51
NT-500M-human	480M	3.19	50B	0 || 0	55.43
NT-500M-1000g	480M	3.19	50B	0 || 1	58.23
NT-2500M-1000g	2537M	19.44	300B	0 || 1	61.41
NT-2500M-multi	2537M	19.44	300B	7 || 9	66.93
DNABERT-2	117M	1.00	262B	8 || 4	66.80
DNABERT-2◆	117M	1.00	263B	**11** || **10**	**67.77**

DNABERT-2 vs DNABERT: 3x faster, unlimited input length, and much better performance.

DNABERT-2 vs NT-2.5B: 19x faster, 92x less training cost, and similarly performance.

DNABERT-2

Results on GUE+

Task	SC (Fungi)	SC (Virus)	EPI (Human)					
Dataset			**0**	**1**	**2**	**3**	**4**	**5**
DNABERT (6-mer)	89.29	44.51	-	-	-	-	-	-
NT-2500M-multi	92.85	45.00	61.91	72.15	73.13	79.49	86.48	68.64
DNABERT-2	**93.04**	**48.50**	**76.21**	**79.19**	**83.50**	**86.71**	**92.90**	**73.70**

DNABERT-2 vs NT-2.5B: better performance on long-sequence tasks

DNABERT-2

Code and Model

The code is publicly available at https://github.com/MAGICS-LAB/DNABERT_2

The trained models is available at HuggingFace:
https://huggingface.co/zhihan1996/DNABERT-2-117M