Artificial Intelligence for Analysis of Nucleic Acid Sequences

Nataliya Logvina

Advisor: Radoslav Neychev

Main idea

Adapter-based multitask fine-tuning application for genome Regulatory Elements prediction.

Questions:

Why do we need predict regulatory elements?

Why we suggest adapters and multitask fine-tuning?

Why predict regulatory elements?

Disease Insight & Treatment Enhance understanding of the disease mechanisms for better diagnostics and targeted treatments.

Precision Medicine Facilitate individualized healthcare by predicting disease susceptibility and treatment response.

Gene Therapy Design safe and effective therapies

Agricultural Sciences Targeted modification of desirable traits.

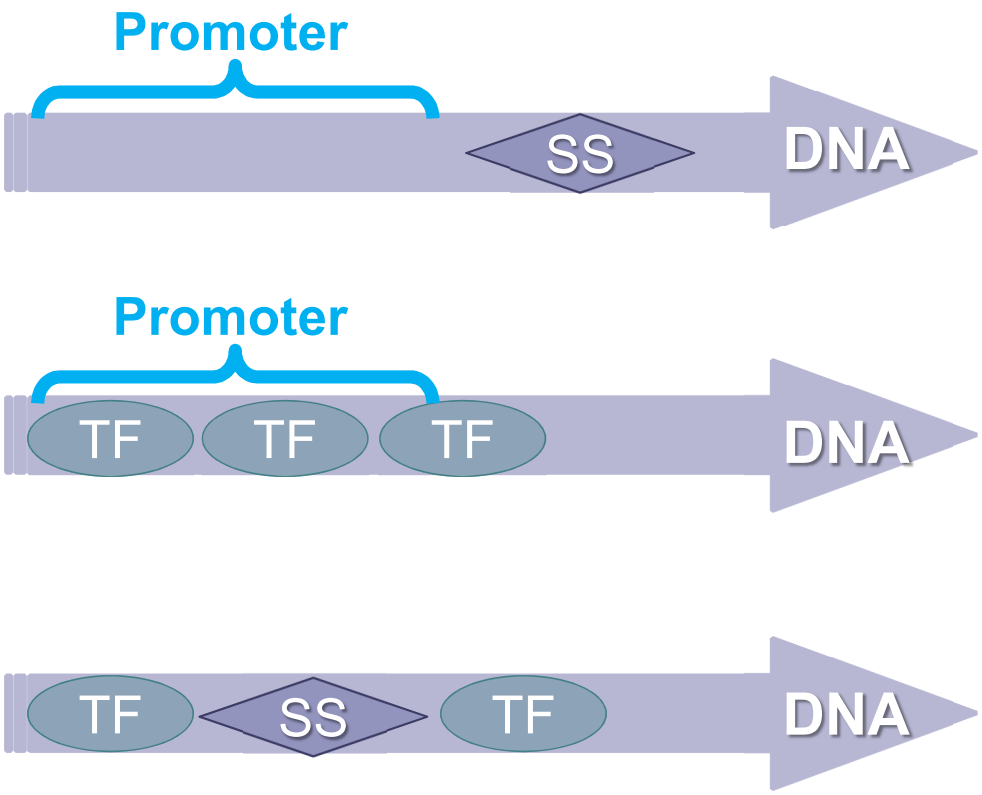
Environmental Genomics Understanding the genetic responses to the environmental changes

Classical tasks for prediction of genomic regulatory elements

- Promoter prediction (PD)
- Transcription Factors binding sites prediction (TF)
- Splice Sites prediction (SSP)
- Epigenetic Marks prediction (EMP)

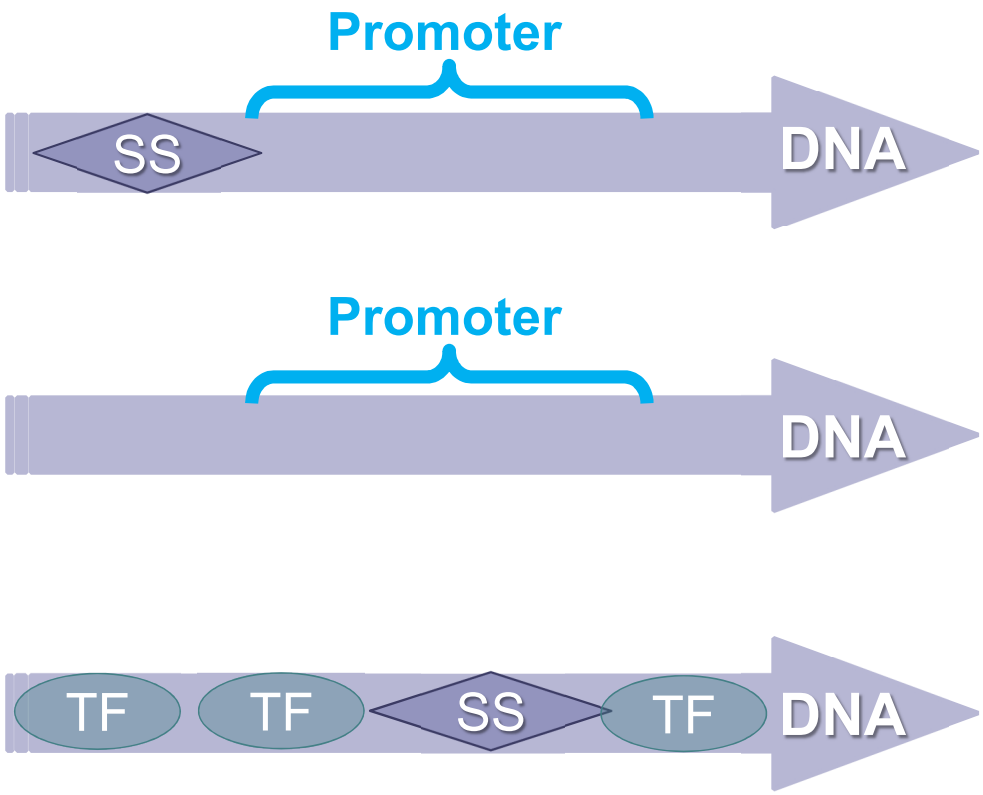
Positional dependencies between the corresponding regulatory elements

Usual arrangement



etc…

Improbable combinations



etc…

Proposed approach

Application of NLP techniques to analyze DNA sequences:

1. Using BERT base model pretrained on human genome with MLM task
2. Add adapters to fine-tune on separate tasks
3. Leverage the adapter Fusion to enhance performance, utilizing the capability to handle multiple tasks

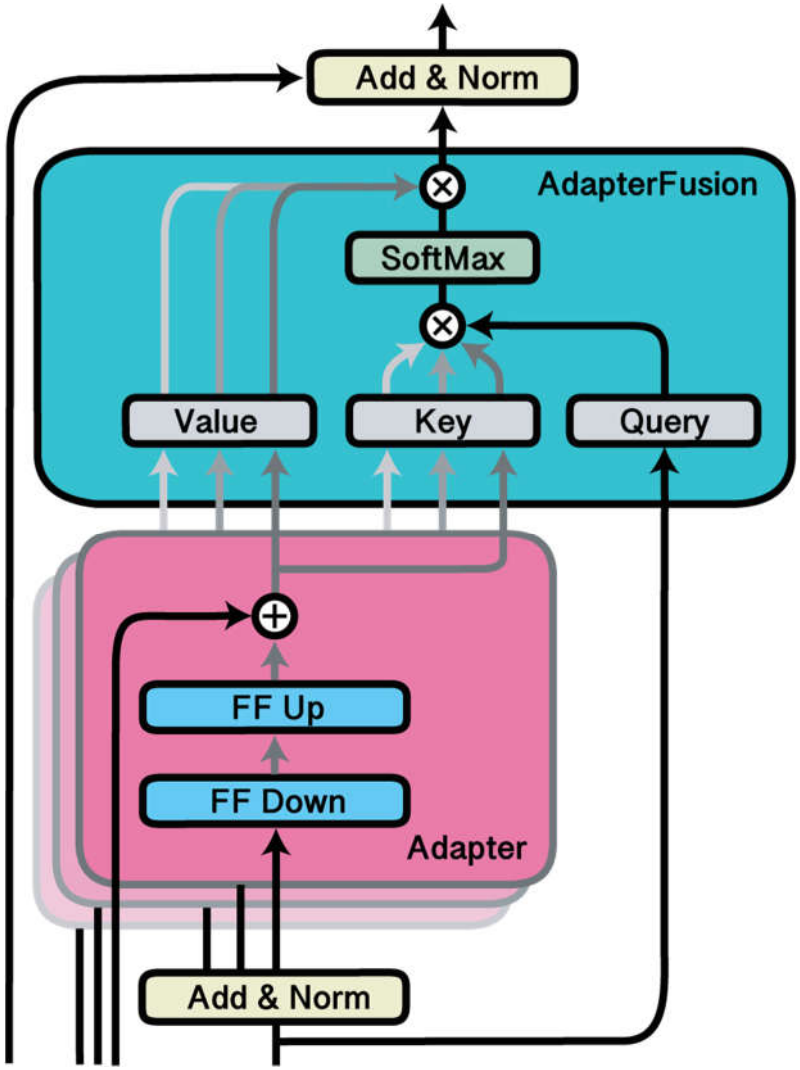
Adapter Fusion

1. The new Attention (Fusion) layer is added on top of a group of adapters

   This layer selects, merges and utilizes the capabilities of different adapters simultaneously, enhancing the overall predictive performance.

2. Advantages over Multitask Fine-Tuning:

   - Moduarity
   - Flexibility
   - No catastrophic forgetting
   - Reduced Overfitting Risk



https://docs.adapterhub.ml/

GUE dataset details

GUE (Genome Understanding Evaluation) is a comprehensive multi-species genome classification dataset, bringing together 28 distinct datasets across 7 tasks, specifically designed to pose a greater challenge for DNA Language Models

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

https://arxiv.org/pdf/2306.15006.pdf

Results comparison

	Yeast	Mouse	Human			
	EMP	TF-M	TF-H	PD	CPD	SSP
DNABERT (3-mer)	49.54	57.73	64.43	84.63	**72.96**	84.14
DNABERT (4-mer)	48.59	59.58	64.41	82.99	71.10	84.05
DNABERT (5-mer)	48.62	54.85	50.46	84.04	72.03	84.02
DNABERT (6-mer)	49.10	56.43	64.17	81.70	71.81	84.07
NT-500M-human	45.35	45.24	50.82	85.51	66.54	79.71
NT-500M-1000g	47.68	49.31	58.92	86.58	69.13	80.97
NT-2500M-1000g	50.86	56.82	61.99	86.61	68.17	85.78
NT-2500M-multi	58.06	67.01	63.32	**88.14**	71.62	**89.36**
DNABERT-2	55.98	67.99	**70.10**	84.21	70.52	84.99
DNABERT-2◆	**58.83**	**71.21**	66.84	83.81	71.07	85.93
Single Adapter	44.21	51.87	62.46	79.71	69.87	84.28
Fusion	49,70	68.46	62.56	78.88	70.02	83.12

Note: DNABERT-2◆ was additionally pretrained on MLM task with GUE dataset

https://arxiv.org/pdf/2306.15006.pdf

Effect of Adapter Fusion on TF-M datasets

Mouse Transcription Factors					
TF	**Single adapter**		**Adapter fusion**		**DS_Length**
	eval	***test***	***eval***	***test***	
0	0.5360	0.5207	0.5336	0.4693	6478
1	0.8054	0.8184	0.8053	0.8185	53952
2	0.8479	0.7866	0.6608	0.6349	2620
3	0.6364	0.6843	0.1733	0.2223	1904
4	0.4749	0.4483	0.4747	0.4485	15064
Avg	0.6601	0.6517	0.52954	0.5187	

Adapter Fusion significantly improves model performance on low resource tasks without negative influence on the high-resource tasks

Effect of Adapter Fusion on yeast datasets

EMP Task	Single Adapter	Adapter Fusion	Delta
H3	71.90	73.30	1.40
H3K14ac	31.91	40.04	8.13
H3K36me3	42.81	47.65	4.84
H3K4me1	36.64	39.19	2.55
H3K4me2	28.29	32.77	4.48
H3K4me3	20.49	44.21	23.72
H3K79me3	59.62	61.48	1.86
H3K9ac	40.94	42.53	1.59
H4	76.52	78.75	2.23
H4ac	32.97	37.12	4.15
Average	44.21	49.70	5.50

Adapter Fusion significantly improves model performance on evolutionary distant dataset/

Conclusions

1. Adapter Fusion significantly boosts performance on low-resource datasets, nearing the results of foundational models.

2. Proposed technique enhances model effectiveness while demanding far fewer computational resources than foundational models.

Suplementary

Tasks and datasets: GUE dataset details

Task	Metric	Datasets	Train / Dev / Test
Core Promoter Detection	**mcc**	tata	4904 / 613 / 613
		notata	42452 / 5307 / 5307
		all	47356 / 5920 / 5920
Promoter Detection	**mcc**	tata	4904 / 613 / 613
		notata	42452 / 5307 / 5307
		all	47356 / 5920 / 5920
Transcription Factor Prediction (Human)	**mcc**	wgEncodeEH000552	32378 / 1000 / 1000
		wgEncodeEH000606	30672 / 1000 / 1000
		wgEncodeEH001546	19000 / 1000 / 1000
		wgEncodeEH001776	27294 / 1000 / 1000
		wgEncodeEH002829	19000 / 1000 / 1000
Splice Site Prediction	**mcc**	reconstructed	36496 / 4562 / 4562
Transcription Factor prediction (Mouse)	**mcc**	Ch12Nrf2Iggrab	6478 / 810 / 810
		Ch12Znf384hpa004051Iggrab	53952 / 6745 / 6745
		MelJundIggrab	2620 / 328 / 328
		MelMafkDm2p5dStd	1904 / 239 / 239
		MelNelfeIggrab	15064 / 1883 / 1883
Epigenetic Marks Prediction	**mcc**	H3	11971 / 1497 / 1497
		H3K14ac	26438 / 3305 / 3305
		H3K36me3	27904 / 3488 / 3488
		H3K4me1	25341 / 3168 / 3168
		H3K4me2	24545 / 3069 / 3069
		H3K4me3	29439 / 3680 / 3680
		H3K79me3	23069 / 2884 / 2884
		H3K9ac	22224 / 2779 / 2779
		H4	11679 / 1461 / 1461
		H4ac	27275 / 3410 / 3410
Virus	**f1**	Covid variant classification	77669 / 7000 / 7000

Recent State-of-the-Art models

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**

Both DNABERT-2 and Nucleotide Transformer (NT) were pretrained on multispecies datasets.

https://doi.org/10.1101/2021.04.27.441365

Results comparison

	Yeast	Mouse	Human			
	EMP	TF-M	TF-H	PD	CPD	SSP
DNABERT (3-mer)	49.54	57.73	64.43	84.63	**72.96**	84.14
DNABERT (4-mer)	48.59	59.58	64.41	82.99	71.10	84.05
DNABERT (5-mer)	48.62	54.85	50.46	84.04	72.03	84.02
DNABERT (6-mer)	49.10	56.43	64.17	81.70	71.81	84.07
NT-500M-human	45.35	45.24	50.82	85.51	66.54	79.71
NT-500M-1000g	47.68	49.31	58.92	86.58	69.13	80.97
Single Adapter	44.21	51.87	62.46	79.71	69.87	**84.28**
Fusion	**49.70**	**68.46**	62.56	78.88	70.02	83.12

Note: DNABERT-2◆ was additionally pretrained on MLM task with GUE dataset

https://arxiv.org/pdf/2306.15006.pdf