DNABERT

Goal of the paper

DNA exhibits:

- polysemy (meaning the same sequence having multiple meanings)
- distant semantic relationships

Therefore DNA is similar to a language.

Can BERT be applied to DNA and replace previous techniques

Introduction

- Evidence of polysemy and distant semantic relationships (CREs)
- CRE, non coding regions of DNA regulating transcription of neighbouring genes. Cis meaning same side.

A successful model should

(1) Take in global contextual information to distinguish polysemous CREs
(2) Be transferable to different tasks
(3) Generalise well using limited data

Previous techniques (RNNs, CNNs) don’t satisfy these requirements

DNABERT Architecture

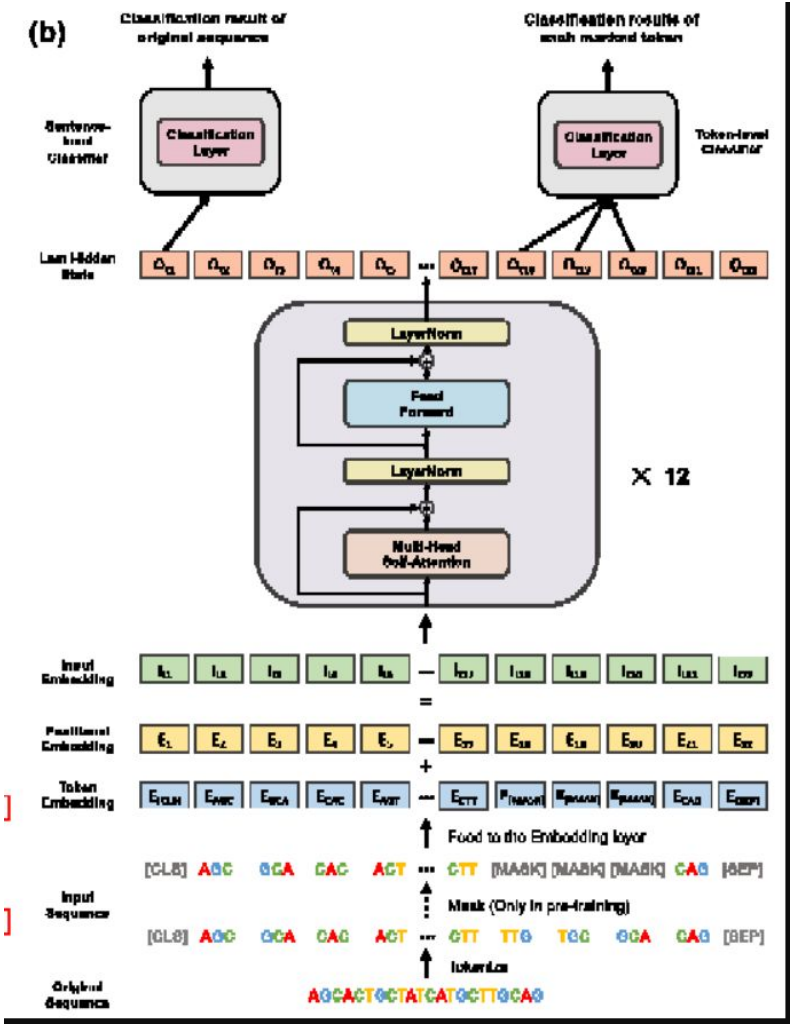
Based on pre-training and fine-tuning paradigm. Consists of 12 encoding layers.

From their Supplementary Materials:

"Starting from the pre-trained parameters, we exploit labeled data to optimize the DNABERT and the output layer simultaneously"

It differs to BERT:

- No NSP
- Adjusting sequence length
- Must predict contiguous k tokens



Downstream tasks

DNABERT was fine-tuned on three separate tasks:

1. Promoter prediction
2. Transcription factor binding site prediction
3. Splice site prediction

Compared DNABERT performance in these tasks against current best models

Pretraining and fine tuning

Contiguous masking (15% at the start increased to 20% near the end of training)

No masking when fine-tuning

Training sequences of 10-510 bps was generated from the human genome via splitting and random sampling

Linear warm-up of learning rate to peak value then linear decay to 0

Results

Promoter prediction

Data from Eukaryotic Promoter Database.

Started with 3,065 human TATA and 26,533 non-TATA promoter-containing sequences ranging from -5,000 to +5,000 bp with +1 being the transcription start site.

Promoter prediction models

Two models were fine-tuned to fairly benchmark it against existing tools

Model 1: DNABERT-Prom-300 uses 300 bp long sequences -249 to +50bp around TSS

First they predicted proximal promoter regions (-249~+50 bp sequences around TSS) then extended it to predicted core promoter regions (-34~+35 bp subsequences of previous sequences around TSS)

Promoter prediction models

Model 2: DNABERT-Prom-scan

Mimics real-world situation of scanning long genomic regions with a sliding window to obtain 1001 bp sequences for promoter prediction.

Used the whole 10,000 bp sequences from the EPD with a step-size of 100.

Highly imbalanced dataset

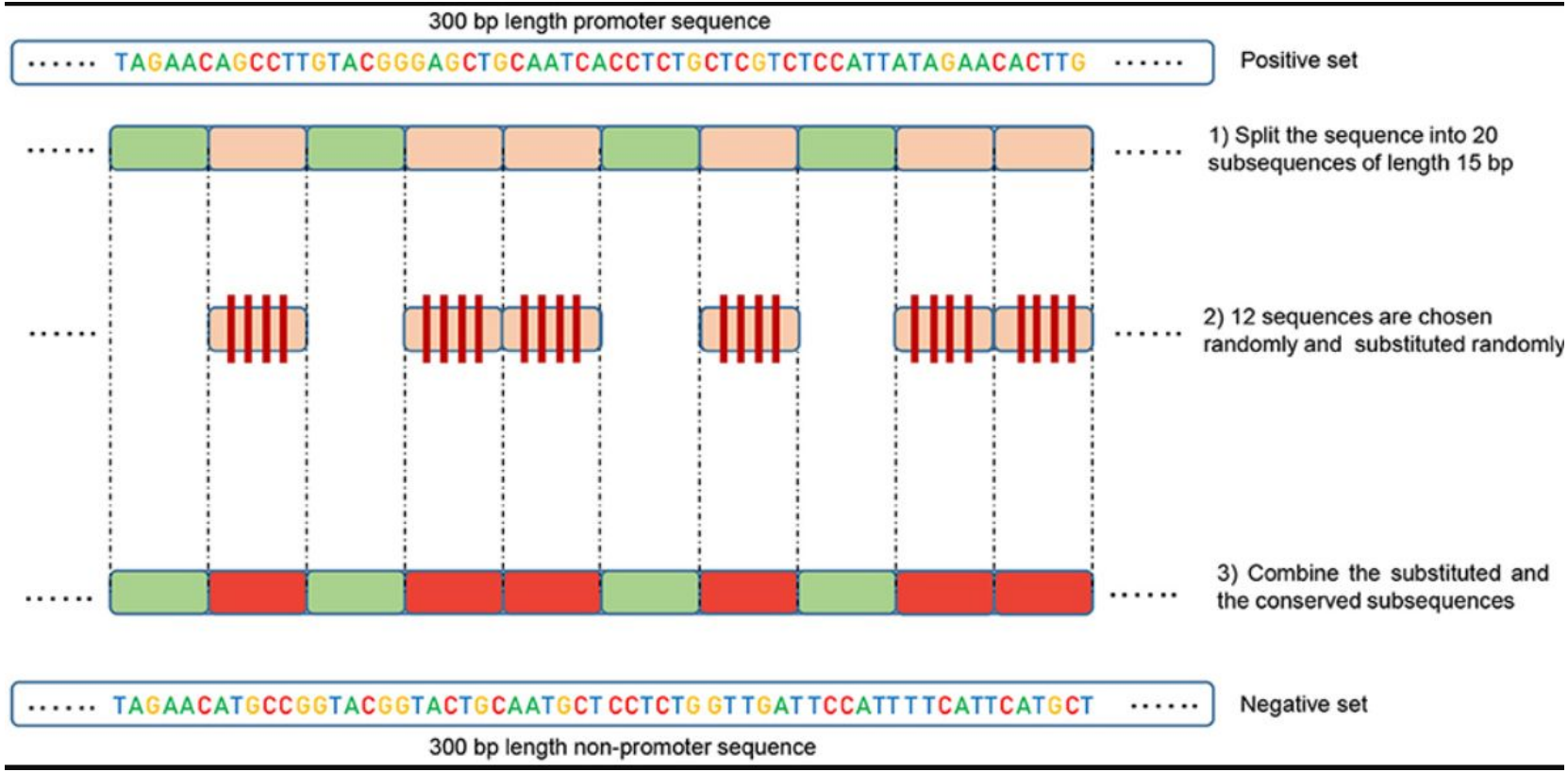
DNA-Prom-300

Positive class: -249 to +50bp around TSS

TATA negative class: select 3,065 300 bp regions not in the -249~+50 range containing the TATA motif. TATA box is also located in the same region as it does in the positive class (25 bp upstream of TSS)

non-TATA negative class: Based on DeePromoter paper. From positive class sequence generate 20 equal length subsequences, for a random 12/20 of them do random substitution. Combine with remaining 8 conserved subsequences to generate full length negative class sequence. This will contain similar motifs to original positive sequence.

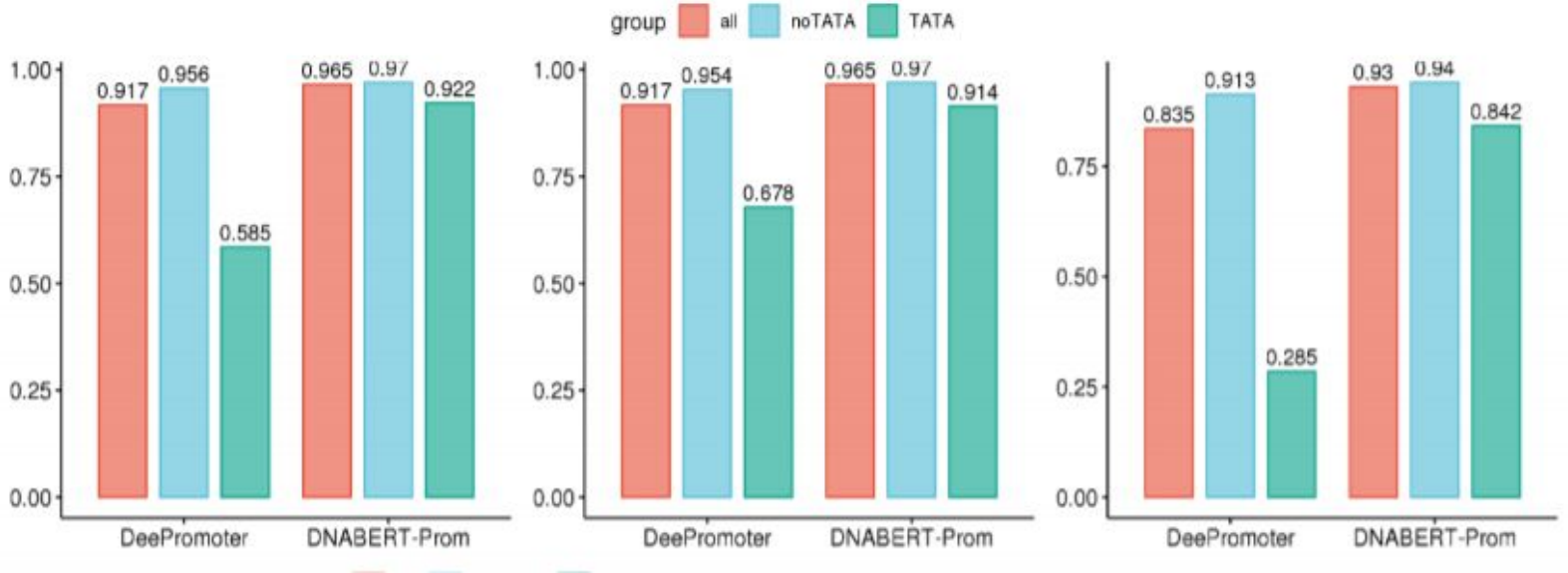
Negative Dataset construction from DeePromoter paper



Accuracy

Proximal promoter prediction (DNABERT-Prom-300) accuracy against DeePromoter (a CNN and LSTM based proximal promoter prediction model)

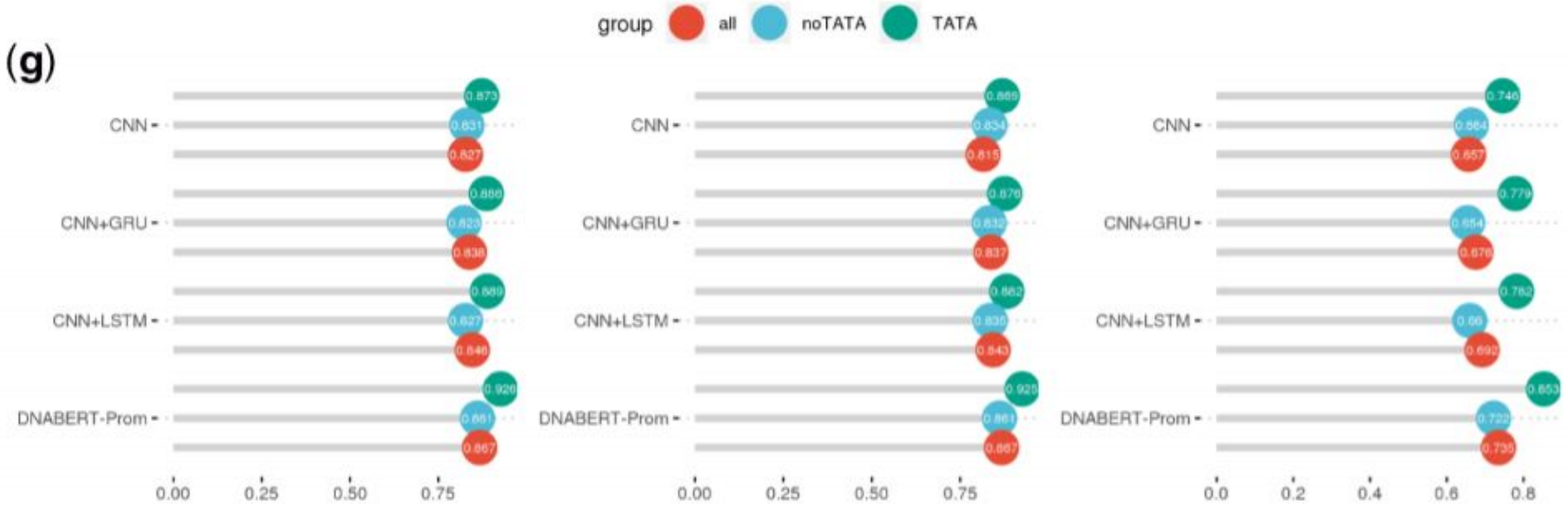
From left to right: accuracy, F1 and MCC



Accuracy

Core promoter prediction accuracy (DNABERT-Prom-core)

Left to right: accuracy, F1 and MCC



Transcription Factor Binding Site prediction

Transcription factors are what start transcription in DNA and they bind to specific regions of non-coding DNA (including promoters).

Task 2 is to predict TFBS in target cis-regulatory regions and curate TF-binding profiles.

DNABERT achieved an accuracy greater than 0.9 which outperforms the current best model.

Visualisation

DNABERT includes a visualisation module which uses attention to aid in interpretability of the trained model’s outputs.

Splice site prediction

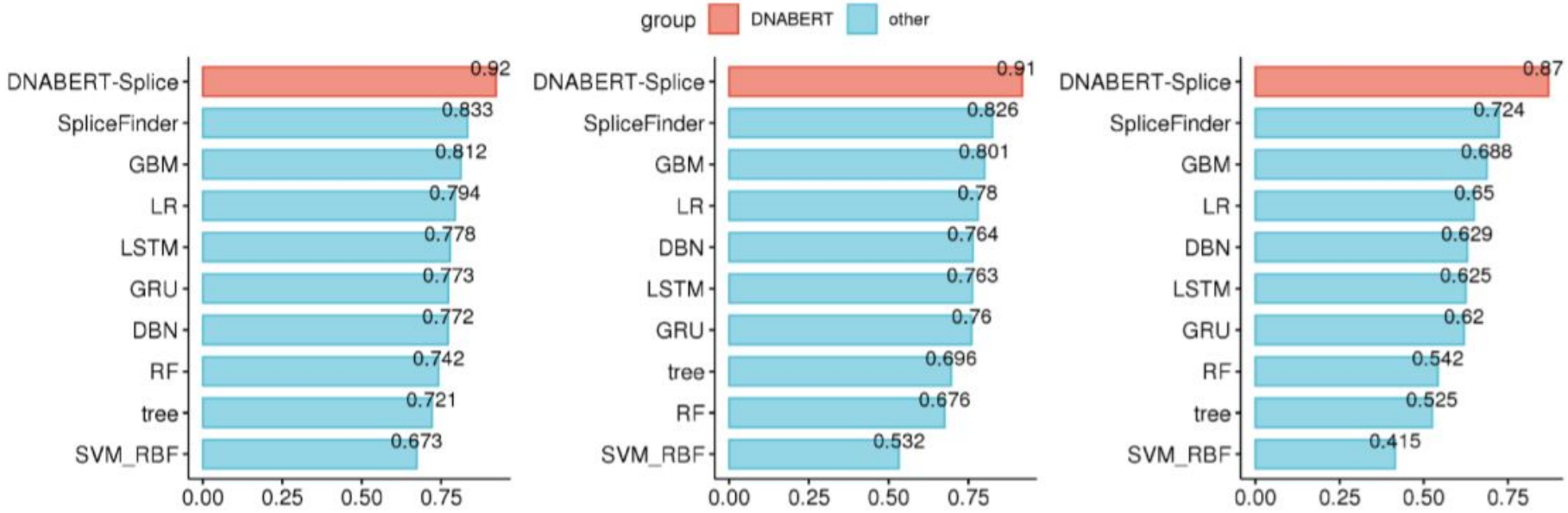
Splice sites in DNA allow the same sequence of DNA to produce different mRNA combinations and therefore different proteins (alternative splicing).

Using 400 bp long sequences predict 3 classes: donor, acceptor and non splice site. Too simple, and does not predict non-canonical splice sites (ones which don’t contain invariant GT and AG dimers often seen in splice sites)

As expected all baseline models and DNABERT perform well for this oversimplified task.

Repeatedly build a new model by predicting on sliding window scans. Only a TP if the splice site is exactly in the center. All false positives are added to the negative set and a new model is trained until the number of false positives is low.

Accuracy, F1 and MCC (harder method)



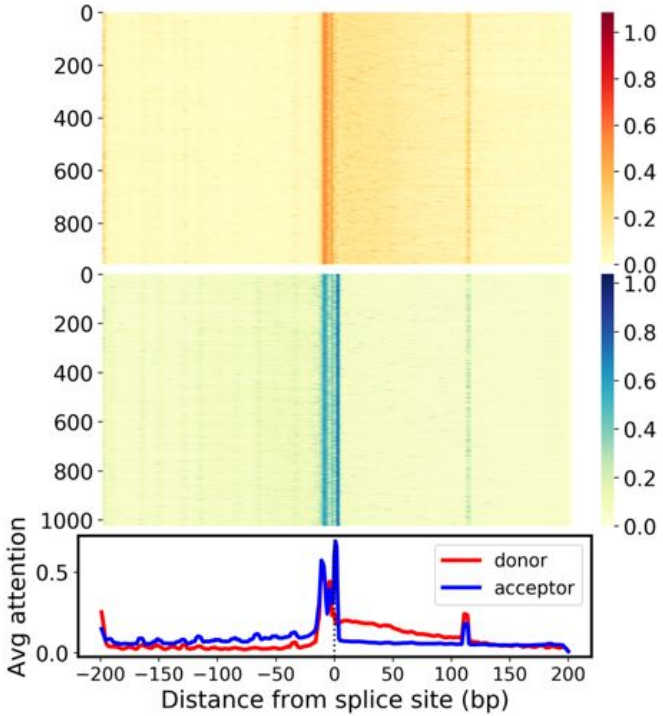


Fig. S12. Attention landscape of splice donor (top) and acceptor (bottom).

Attention landscape revealed high attention upon intronic regions (regions in between exons) highlighting the presence and functional importance of various intronic splicing enhancers and silencers acting as CREs for splicing

Identifying functional variants

Functional variants -> slightly mutated sequences

Genetic effect of this mutation can be calculated as a change in prediction probability. Mutation with large changes in scores were queried in certain databases containing certain clinical and functional variants.

DNABERT shows attention at/around variant of interest.

Discussion

- Pretraining significantly improves performance
- DNABERT has transferable knowledge to other organisms and genomes