Bioinformatics, 37(15), 2021, 2112–2120
doi: 10.1093/bioinformatics/btab083
Advance Access Publication Date: 4 February 2021
Original Paper

Genome analysis

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

Yanrong Ji[1,†], Zhihan Zhou[2,†], Han Liu[2,*] and Ramana V. Davuluri[3,*]

[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA, [2]Department of Computer Science, Northwestern University, Evanston, IL 60208, USA and [3]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Dr. Janet Kelso

Received on September 10, 2020; revised on December 31, 2020; editorial decision on January 25, 2021; accepted on February 1, 2021

Abstract

Motivation: Deciphering the language of non-coding DNA is one of the fundamental problems in genome research. Gene regulatory code is highly complex due to the existence of polysemy and distant semantic relationship, which previous informatics methods often fail to capture especially in data-scarce scenarios.

Results: To address this challenge, we developed a novel pre-trained bidirectional encoder representation, named DNABERT, to capture global and transferrable understanding of genomic DNA sequences based on up and down-stream nucleotide contexts. We compared DNABERT to the most widely used programs for genome-wide regulatory elements prediction and demonstrate its ease of use, accuracy and efficiency. We show that the single pre-trained transformers model can simultaneously achieve state-of-the-art performance on prediction of promoters, splice sites and transcription factor binding sites, after easy fine-tuning using small task-specific labeled data. Further, DNABERT enables direct visualization of nucleotide-level importance and semantic relationship within input sequences for better interpretability and accurate identification of conserved sequence motifs and functional genetic variant candidates. Finally, we demonstrate that pre-trained DNABERT with human genome can even be readily applied to other organisms with exceptional performance. We anticipate that the pre-trained DNABERT model can be fined tuned to many other sequence analyses tasks.

Availability and implementation: The source code, pretrained and finetuned model for DNABERT are available at GitHub (https://github.com/jerryji1993/DNABERT).

Contact: ramana.davuluri@stonybrookmedicine.edu or hanliu@northwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Deciphering the language of DNA for hidden instructions has been one of the major goals of biological research (Andersson and Sandelin, 2020). While the genetic code explaining how DNA is translated into proteins is universal, the regulatory code that determines when and how the genes are expressed varies across different cell-types and organisms (Nirenberg *et al.*, 1965). Same *cis*-regulatory elements (CREs) often have distinct functions and activities in different biological contexts, while widely spaced multiple CREs may cooperate, resulting in context-dependent use of alternative promoters with varied functional roles (Davuluri *et al.*, 2008; Gibcus and Dekker, 2012; Ji *et al.*, 2020; Vitting-Seerup and Sandelin, 2017). Such observations suggest existence of polysemy and distant semantic relationship within sequence codes, which are key properties of natural language. Previous linguistics studies confirmed that the DNA, especially the non-coding region, indeed exhibits great similarity to human language, ranging from alphabets and lexicons to grammar and phonetics (Brendel and Busse, 1984; Head, 1987; Ji, 1999; Mantegna *et al.*, 1994; Searls, 1992; 2002). However, how the semantics (i.e. functions) of CREs vary across different contexts (up and downstream nucleotide sequences) remains largely unknown.

In recent years, many computational tools have been developed by successfully applying deep learning techniques on genomic sequence data to study the individual aspects of *cis*-regulatory

landscapes, including DNA-protein interactions (Alipanahi *et al.*, 2015), chromatin accessibility (Kelley *et al.*, 2016), non-coding variants (Zhou and Troyanskaya, 2015) and others. Most methods adopted Convolutional Neural Network (CNN)-based architecture (Zou *et al.*, 2019). Other tools focus on the sequential characteristic of DNA and attempt to capture the dependency between states by applying Recurrent Neural Network (RNN)-based models, such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho *et al.*, 2014) networks. Several hybrid methods were also proposed to integrate the advantages of the two model architectures (Hassanzadeh and Wang, 2016; Quang and Xie, 2016; Shen *et al.*, 2018).

To better model DNA as a language, an ideal computational method should (i) globally take all the contextual information into account to distinguish polysemous CREs; (ii) develop generic understanding transferable to various tasks; (iii) generalize well when labeled data is limited. However, both CNN and RNN architectures fail to satisfy these requirements (Fig. 1a) (Bengio *et al.*, 2013; LeCun *et al.*, 2015). CNN is usually unable to capture semantic dependency within long-range contexts, as its capability to extract local features is limited by the filter size. RNN models (LSTM, GRU), although able to learn long-term dependency, greatly suffer from vanishing gradient and low-efficiency problem when it sequentially processes all past states and compresses contextual information into a bottleneck with long input sequences. In addition, most existing models require massive amount of labeled data, resulting in limited performance and applicability in data-scarce scenarios, where high quality data with labels is expensive and time-consuming to obtain.

To address the above limitations, we adapted the idea of Bidirectional Encoder Representations from Transformers (BERT) model (Devlin *et al.*, 2018) to genomic DNA setting and developed a deep learning method called DNABERT. DNABERT applies Transformer, an attention-based architecture that has achieved state-of-the-art performance in most natural language processing tasks (Vaswani *et al.*, 2017). We demonstrate that DNABERT resolves the above challenges by (i) developing general and transferable understandings of DNA from the purely unlabeled human genome, and utilizing them to generically solve various sequence-related tasks in a 'one-model-does-it-all' fashion; (ii) globally capturing contextual information from the entire input sequence with attention mechanism; (iii) achieving great performance in data-scarce scenarios; (iv) uncovering important subregions and potential relationships between different *cis*-elements of a DNA sequence, without any human guidance; (v) successfully working in a cross-organism manner. Since the pre-training of DNABERT model is resource-intensive (about 25 days on 8 NVIDIA 2080Ti GPUs), as a major contribution of this study, we provide the source code and pretrained model on GitHub for future academic research.

2 Materials and methods

2.1 The DNABERT model

BERT is a transformer-based contextualized language representation model that has achieved superhuman performance in many natural language processing (NLP) tasks. It introduces a paradigm of pre-training and fine-tuning, which first develops general-purpose understandings from massive amount of unlabeled data and then solves various applications with task-specific data with minimal architectural modification. DNABERT follows the same training process as BERT. More details are included in Supplementary Material.

DNABERT first takes a set of sequences represented as k-mer tokens as input (Fig. 1b). Each sequence is represented as a matrix M by embedding each token into a numerical vector. Formally,

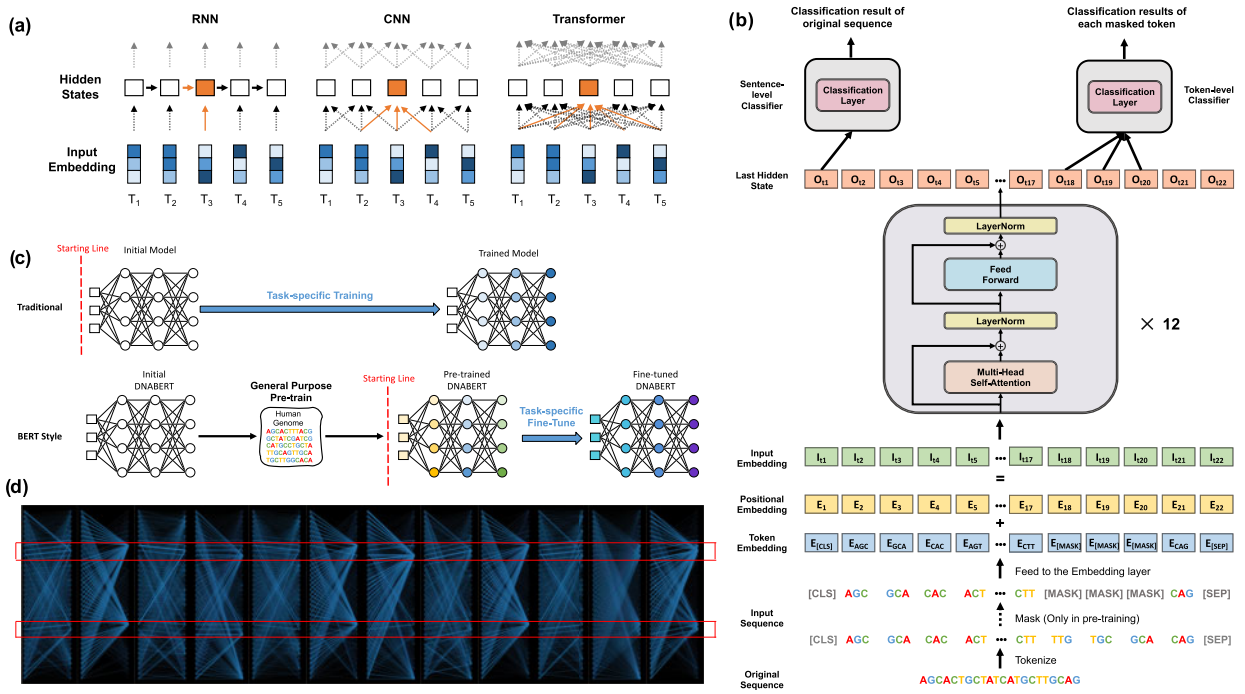

Fig. 1. Details of architecture and characteristics of DNABERT model.(**a**) Differences between RNN, CNN and Transformer in understanding contexts. T1 to 5 denotes embedded tokens which were input into models to develop hidden states (white boxes, orange box is the current token of interest). RNN propagates information through all hidden states, and CNN takes local information in developing each representation. In contrast, Transformers develop global contextual embedding via self-attention. (**b**) DNABERT uses tokenized k-mer sequences as input, which also contains a CLS token (a tag representing meaning of entire sentence), a SEP token (sentence separator) and MASK tokens (to represent masked k-mers in pre-training). The input passes an embedding layer and is fed to 12 Transformer blocks. The first output among last hidden states will be used for sentence-level classification while outputs for individual masked token used for token-level classification. Et, It and Ot denote the positional, input embedding and last hidden state at token t, respectively. (**c**) DNABERT adopts general-purpose pre-training which can then be fine-tuned for multiple purposes using various task-specific data. (d) Example overview of global attention patterns across 12 attention heads showing DNABERT correctly focusing on two important regions corresponding to known binding sites within sequence (boxed regions, where self-attention converged)

DNABERT captures contextual information by performing the multi-head self-attention mechanism on M:

$$MultiHead(M) = Concat(head_1, \ldots, head_h)W^O \quad (1)$$

where

$$head_i = softmax\left(\frac{MW_i^Q MW_i^K T}{\sqrt{d_k}}\right) \cdot MW_i^V \quad (2)$$

W^O and $W_i^Q, W_i^K, W_i^V \left\{W_i^Q, W_i^K, W_i^V\right\}_{i=0}^{h}$ are learned parameters for linear projection. *head* calculates the next hidden states of M by first computing the attentions scores between every two tokens and then utilizing them as weights to sum up lines in MW_i^V. *MultiHead*() concatenates results of h independent *head* with different set of $\{W_i^Q, W_i^K, W_i^V\}$. The entire procedure is performed L times with L being number of layers.

Similar to BERT, DNABERT also adopts pre-training—fine-tuning scheme (Fig. 1c). However, we significantly modified the pre-training process from the original BERT implementation by removing next sentence prediction, adjusting the sequence length and forcing the model to predict contiguous k tokens adapting to DNA scenario. During pre-training, DNABERT learns basic syntax and semantics of DNA via self-supervision, based on 10 to 510-length sequences extracted from human genome via truncation and sampling. For each sequence, we randomly mask regions of k contiguous tokens that constitute 15% of the sequence and let DNABERT to predict the masked sequences based on the remainder, ensuring ample training examples. We pre-trained DNABERT with cross-entropy loss: $L = \sum_{i=0}^{N} -y'_i \log(y_i)$. Here, y'_i and y_i are the ground-truth and predicted probability for each of N classes. The pre-trained DNABERT model can be fine-tuned with task-specific training data for applications in various sequence- and token-level prediction tasks. We fine-tuned DNABERT model on three specific applications—prediction of promoters, transcription factor binding sites (TFBSs) and splice sites—and benchmarked the trained models with the current state-of-the-art tools.

2.2 Training of the DNABERT model

2.2.1 Tokenization

Instead of regarding each base as a single token, we tokenized a DNA sequence with the k-mer representation, an approach that has been widely used in analyzing DNA sequences. The k-mer representation incorporates richer contextual information for each deoxynucleotide base by concatenating it with its following ones. The concatenation of them is called a k-mer. For example, a DNA sequence 'ATGGCT' can be tokenized to a sequence of four 3-mers: {ATG, TGG, GGC, GCT} or to a sequence of two 5-mers: {ATGGC, TGGCT}. Since different k leads to different tokenization of a DNA sequence. In our experiments, we respectively set k as 3,4,5 and 6 and train 4 different models: DNABERT-3, DNABERT-4, DNABERT-5, DNABERT-6. For DNABERT-k, the vocabulary of it consists of all the permutations of the k-mer as well as 5 special tokens: [CLS] stands for classification token; [PAD] stands for padding token, [UNK] stands for unknown token, [SEP] stands for separation token and [MASK] stands for masked token. Thus, there are $4^k + 5$ tokens in the vocabulary of DNABERT-k.

2.2.2 Pre-training

Following previous works (Devlin *et al.*, 2018; Liu *et al.*, 2019; Yang *et al.*, 2019), DNABERT takes a sequence with a max length of 512 as input. As illustrated in Figure 1b, for a DNA sequence, we tokenized it into a sequence of k-mers and added a special token [CLS] at the beginning of it (which represents the whole sequence) as well as a special token [SEP] at the end (which denotes the end of sequence). In the pre-training step, we masked contiguous k-length spans of certain k-mers, considering a token could be trivially inferred from the immediately surrounding k-mers (total ~15% of input sequence), while in the fine-tuning, we skipped the masking step and directly fed the tokenized sequence to the Embedding layer.

We generated training data from human genome via two approaches: direct non-overlap splitting and random sampling, with length of the sequence between 5 and 510. We pre-trained DNABERT for 120k steps with a batch size of 2000. In this first 100k steps, we masked 15 percent of k-mers in each sequence. In the last 20k steps, we increased the masking rate to 20 percent. The learning rate was linearly increased (i.e. warm-up) from 0 to 4e-4 in the first 10k steps and then linearly decreased to 0 after 200k steps (Supplementary Fig. S1). We stopped the training procedure after 120k steps since we found the loss curve show a sign of plateauing. We used the same model architecture as the BERT base, which consists of 12 Transformer layers with 768 hidden units and 12 attention heads in each layer, and the same parameter setting across all the four DNABERT models during pre-training. We trained each DNABERT model with mixed precision floating point arithmetic on machines with 8 Nvidia2080Ti GPUs.

2.2.3 Fine-tuning

For each downstream application, we started from the pre-trained parameters and fine-tuned DNABERT with task-specific data. We utilized the same training tricks across all the applications, where the learning rate was first linear warmed-up to the peak value and then linear decayed to near 0. We utilized AdamW with fixed weight decay as optimizer and employed dropout to the output layer. We split training data into training set and developing set for hyperparameter tuning. For DNABERT with different k, we slightly adjusted the peak learning rate. The detailed hyperparameter settings were listed in Supplementary Table S5. For sequences longer than 512, we split them into pieces and concatenate their representations as the final representation. This allows DNABERT to process extra-long sequences (DNABERT-XL). DNABERT with k = 3, 4, 5, 6 achieved very similar performances with slight fluctuations. In all experiments, we report results of kmer = 6 since it achieves the best performance.

3 Results

3.1 DNABERT-Prom effectively predicts proximal and core promoter regions

Predicting gene promoters is one of the most challenging bioinformatics problems. We began by evaluating our pre-trained model on identifying proximal promoter regions. To fairly compare with existing tools with different sequence length settings, we fine-tuned two models, named DNABERT-Prom-300 and DNABERT-Prom-scan, using human TATA and non-TATA promoters of 10 000 bp length, from Eukaryotic Promoter Database (EPDnew) (Dreos *et al.*, 2013). We compared DNABERT-Prom-300 with DeePromoter(Oubounyt *et al.*, 2019) using -249 to 50 bp sequences around TSS as positive examples, randomly selected 300 bp-long, TATA-containing sequences as TATA negative examples, and dinucleotide-shuffled sequences as non-TATA negative examples (Supplementary Methods). We compared DNABERT-Prom-scan with currently accessible methods, including recent state-of-the-art methods PromID(Umarov *et al.*, 2019), FPROM (Solovyev *et al.*, 2006), and our previous software FirstEF(Davuluri, 2003), using sliding window-based scans from 10 000 bp-long sequences. To appropriately benchmark with PromID under same setting, we used 1001 bp-long scans, which exceed the length capacity of traditional BERT model. Hence, we developed DNABERT-XL specifically for this task (Supplementary Methods). We used same evaluation criteria used in PromID by scanning sequences and overlapping predictions with -500 to +500 bp of known TSS. The resulting 1001 bp sequences with $\geq$ 50% overlap to -500 to +500 bp of TSS were deemed as positives and the remaining as negatives. For PromID and FPROM, the test set was directly input for evaluation. In contrast, FirstEF first generates genome-wide predictions, which were then aligned to the positive sequences.

DNABERT-Prom outperformed all other models by significantly improved accuracy metrics regardless of different settings (Fig. 2). Specifically, for prom-300 setting TATA promoters, DNABERT-Prom-300 exceeded DeePromoter in accuracy and MCC metrics by

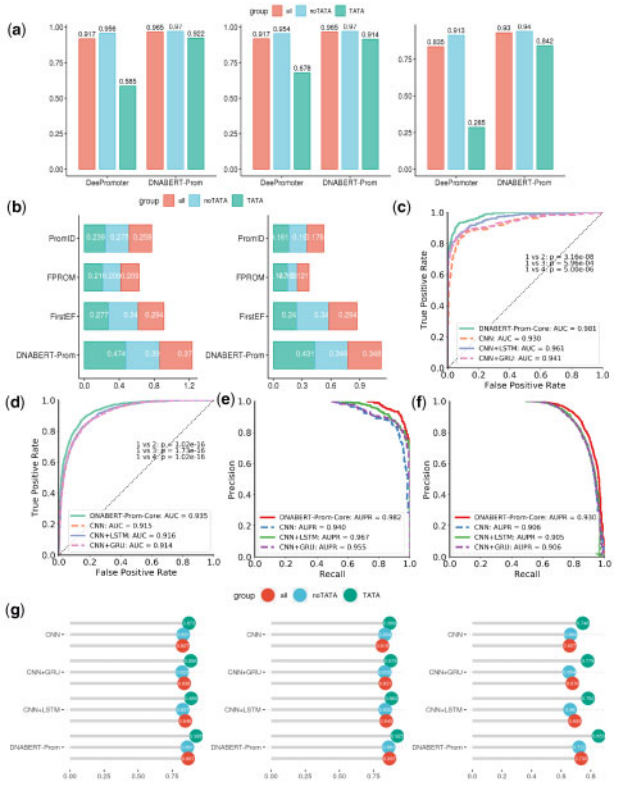

Fig. 2. DNABERT significantly outperforms other models in identifying promoter regions. (**a**) (Left to right) accuracy, F1 and MCC of prom-300 prediction in TATA, no-TATA and combined datasets. (**b**) Stacked barplot showing F1 (left) and MCC (right) of Prom-scan predictions in different settings. (**c–f**) ROC (c, TATA; d, noTATA) and Preci- sion-recall (PR) curves (e, TATA; f, noTATA) with adjusted *P*-values from Delong test. (g) (Left to right) accuracy, F1 and MCC of core promoters prediction in TATA, no-TATA and combined datasets

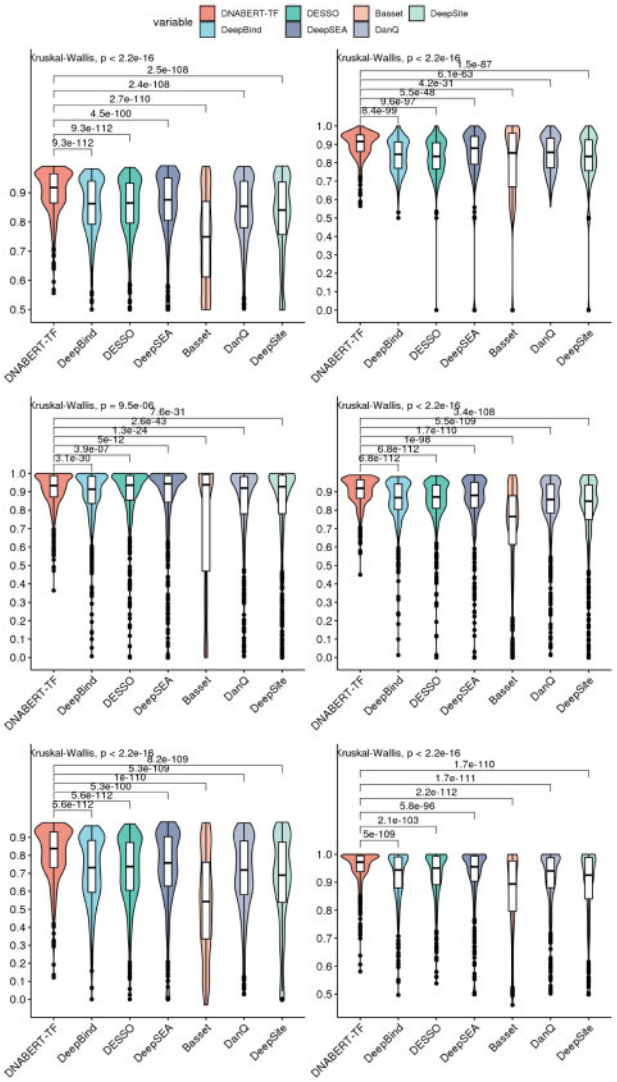

Fig. 3. DNABERT accurately identifies TFBSs.Violin plots showing accuracy (top left), precision (top right), recall (middle left), F1 (middle right), MCC (bottom left)- and AUC (bottom right) of TFBS prediction with ENCODE 690 ChIP-Seq datasets. Pairwise comparison using Wilcoxon one-sided signed-rank test ($n = 690$) and adjusted *P*-values using Benjamini-Hochberg procedure were shown. Global hypothesis testing across all models done by Kruskal-Wallis test ($n = 690$)

0.335 and 0.554, respectively (Fig. 2a). Similarly, we observed significantly improved performance of DNABERT-Prom in both non-TATA and combined cases (Supplementary Fig. S2). Meanwhile, the prom-scan setting is intrinsically more difficult as the classes are highly imbalanced, so all the tested baseline models performed poorly. Among the baselines, FirstEF achieved the best performance with an F1-score of 0.277 for TATA, 0.377 for non-TATA and 0.331 for combined datasets (Fig. 2b). However, DNABERT-Prom-scan achieved F1-score and MCC that largely surpassed FirstEF. Next, we evaluated our model's predictive performance on core promoters, a more challenging problem due to reduced size of the sequence context. We used 70 bp, centered around TSS, of the Prom-300 data and compared with CNN, CNN+LSTM and CNN+GRU. DNABERT-Prom-core clearly outperformed all the three baselines across different datasets (Fig. 2c–g), clearly demonstrating that DNABERT can be reliably fine-tuned to accurately predict both the long proximal promoters and shorter core promoters, relying only on nearby sequence patterns around the TSS region. To further demonstrates the effectiveness of DNABERT-XL, we also conducted experiments on 301 bp-long sequences and 2001 bp-long sequences. Experiments shows that the model achieves a better performance in predicting 2001 bp-long sequences (Supplementary Table S7).

3.2 DNABERT-TF accurately identifies transcription factor binding sites

NextGen sequencing (NGS) technologies have facilitated genome-wide identification of gene regulatory regions in an unprecedented way and unveiled the complexity of gene regulation. An important step in the analyses of in vivo genome-wide binding interaction data is prediction of TFBS in the target cis-regulatory regions and curation of the resulting TF binding profiles. We thus fine-tuned DNABERT-TF model to predict TFBSs in the ChIP-seq enriched regions, using 690 TF ChIP-seq uniform peak profiles from ENCODE database (Dunham *et al.*, 2012) and compared with well-known and previous published TFBS prediction tools, including DeepBind(Alipanahi *et al.*, 2015), DeepSEA(Zhou and Troyanskaya, 2015), Basset (Kelley *et al.*, 2016), DeepSite(Zhang *et al.*, 2020), DanQ(Quang and Xie, 2016) and DESSO (Khamis *et al.*, 2018). DNABERT-TF is the only method with both mean and median accuracy and F1-score above 0.9 (Fig. 3, 0.918 and 0.919), greatly exceeding the second best competitor (DeepSEA, Wilcoxon one-sided signed-rank test, n = 690, adjusted $P = 4.5 \times 10^{-100}$ and 1×10^{-98} for mean). Other tools made many false positive (FP) and false negative (FN) predictions in certain experiments, resulting in even less satisfactory performance, when comparing the mean due to skewed distribution (Supplementary Table S1). Several tools achieved comparable performance with DNABERT in finding the true negatives (TN) for experiments using high-quality data, yet performed poorly when predicting on low-quality experimental data. In contrast, even on low-quality data, DNABERT achieved significantly higher recall than other tools (Fig. 3, middle left). Meanwhile, DNABERT-TF made much fewer FP predictions than any other

model regardless of the quality of the experiment (Fig. 3, top right). These results are further supported by benchmarking using a subset of ChIP-seq profiles with limited number of peaks, where DNABERT-TF consistently outperformed other methods (Supplementary Fig. S3).

To evaluate whether our method can effectively distinguish polysemouscis-regulatory elements, we focused on p53 family proteins (which recognize same motifs) and investigated contextual differences in binding specificities between TAp73-alpha and TAp73-beta isoforms. We overlapped p53, TAp73-alpha and TAp73-beta ChIP-seq peaks from Gene Expression Omnibus (GEO) dataset GSE15780 with binding sites predicted by our P53Scan program (Koeppel *et al.*, 2011; Yoon *et al.*, 2002) and used the resulting ChIP-seq-characterized BS (∼35 bp) to fine-tune our model. DNABERT-TF achieved near perfect performances (∼0.99) on binary classification of individual TFs (Supplementary Table S2). Using input sequences with a much wider context (500 bp), DNABERT-TF effectively distinguished the two TAp73 isoforms with an accuracy of 0.828 (Supplementary Table S2). In summary, DNABERT-TF can accurately identify even very similar TFBSs based on the distinct context windows.

3.3 DNABERT-viz allows visualization of important regions, contexts and sequence motifs

To overcome the common 'black-box' problem, deep learning models need to maintain interpretability, while exceling in performance in comparison to traditional methods. Therefore, to summarize and understand important sequence features on which fine-tuned DNABERT models base classification decisions on, we developed DNABERT-viz module for direct visualization of important regions contributing to the model decision. We demonstrate that DNABERT is naturally suitable for finding important patterns in DNA sequences and understanding their relationship within contexts due to the attention mechanism, thus ensuring model interpretability.

Figure 4a shows the learned attention maps of three TAp73-beta response elements, where DNABERT-viz accurately determines both positions and scores of TFBS predicted by P53Scan in an unsupervised manner. We then aggregated all heatmaps to produce attention landscapes on test sets of Prom-300 and ENCODE 690 TF. For TATA promoters, DNABERT consistently put high attention upon -20 to -30 bp region upstream of TSS where TATA box is located, while for majority of non-TATA promoters a more scattered attention pattern is observed (Fig. 4b). Such pattern is also seen in TF-690 datasets, where each peak displays a distinct set of high attention regions, most of which scattered around the center of the peaks (Supplementary Fig. S4). We specifically looked into examples of individual ChIP-seq experiments to better understand the attention patterns. Most high-quality experiments show enrichment of attention either around the center of the ChIP-seq peaks or on TFBS region (Fig. 4c and Supplementary Fig. S5). In contrast, low-quality ones tend to have dispersed attention without strongly observable pattern, except the high attention only at the beginning of sequences, which is likely due to model bias (Fig. 4d).

We next extended DNABERT-viz to allow for direct visualization of contextual relationship within any input sequence (Fig. 4e). For example, the leftmost plot shows global self-attention pattern of an input sequence in the p53 dataset, where individual attentions from most k-mer tokens over all heads correctly converge at the two centers of the dimeric BS. We can further infer the interdependent relationship between the BS with other regions of input sequence by observing which tokens specifically paid high attention to the site (Fig. 4e, right). Among attention heads, the orange one clearly discovered hidden semantic relationship within context, as it broadly highlights various short regions contributing to attention of this important token CTT. Moreover, three heads (green, purple and pink) successfully relate this token with the downstream half of the dimeric BS, demonstrating contextual understanding of the input sequence.

To extract conserved motif patterns across many input sequences, we applied DNABERT-viz to find contiguous high-attention regions and filtered them by hypergeometric test (Supplementary Methods). The resulting significant motif instances were then aligned and merged to produce position-weight matrices (PWMs). By applying TOMTOM program (Gupta *et al.*, 2007) on the discovered motifs from ENCODE 690 dataset and compared with JASPAR 2018 database, we found that 1595 out of 1999 motifs discovered successfully aligned to validated motifs (Supplementary Fig. S6, q-value < 0.01). Motifs identified are overall of very high quality illustrated by strong similarity to the documented motifs (Supplementary Fig. S7).

We finally applied DNABERT-viz to understand important factors in distinguishing binding sites of TAp73-alpha from beta isoforms. The attention landscape indeed shows many short regions differentially enriched between two isoforms, with alpha having higher attention concentrated at center and beta more scattered into the contexts (Supplementary Fig. S8). Many strong motif patterns extracted were not aligned to JASPAR database except for a few highlighting unknown relationship (Supplementary Fig. S9). Importantly, differential crosstalk between c-Fos, c-Jun and TAp73-alpha/beta isoforms contributes to apoptosis balance (Koeppel *et al.*,

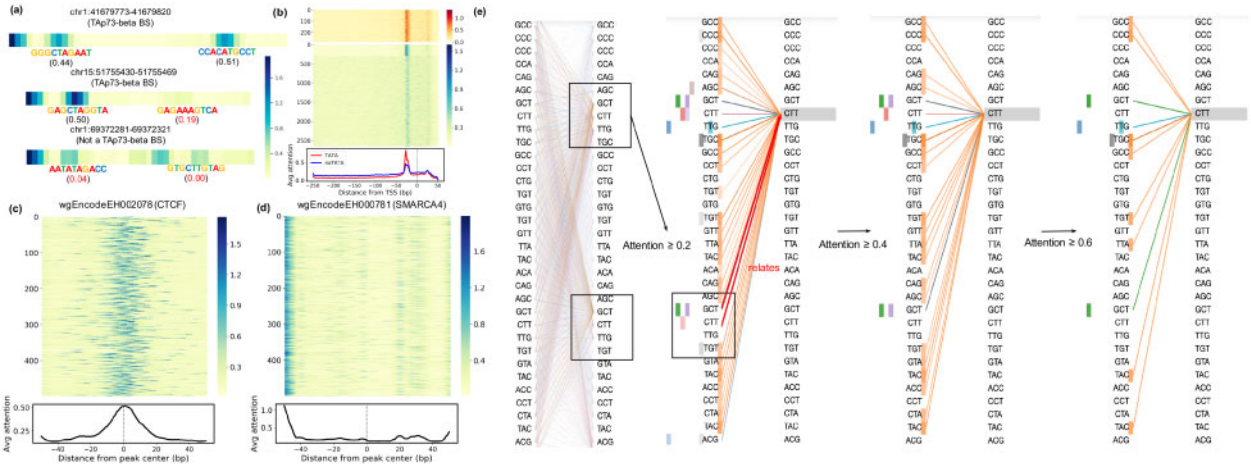

Fig. 4. Visualizations of attention and context by DNABERT-viz.(**a**) Attention maps of two example ChIP-Seq-validated TAp73-beta binding sites (top, middle) and one non-binding site (bottom). Numbers below represent binding scores previously predicted by P53Scan. (**b**) Attention landscapes of TATA (top) and noTATA (bottom) promoters in Prom-300 test set. (**c,d**) Example attention landscapes for individual ENCODE 690 dataset. CTCF (left) is of good quality while SMARCA4 (right) has concerned quality. (**e**) Attention-head (context) plots of a p53 binding site. (left) sentence-level self-attention across all heads; (middle left, middle right, right) attention of the 'CTT' token within one of the important regions, with only attention ≥ 0.2, 0.4 and 0.6 shown respectively. Heatmap on the left shows the corresponding attention head

2011), and DNABERT-viz successfully captured this relationship. To conclude, DNABERT can attain comparable interpretability as CNN-based models in a more straightforward way while greatly surpassing them in prediction performance.

3.4 DNABERT-Splice accurately recognizes canonical and non-canonical splice sites

Predicting splice sites is essential for revealing gene structure and understanding alternative splicing mechanisms. Nevertheless, the presence of both GT-AG-containing non-splice site sequences, and non-canonical splice sites without the dinucleotides, creates difficulty for accurate identification (Wang *et al.*, 2019). Recently, SpliceFinder(Wang *et al.*, 2019) successfully addressed this issue by reconstructing a dataset via recursive inclusion of previously misclassified false positive sequences. To compare with SpliceFinder performance on identical benchmark data, we iteratively rebuilt the same dataset with donor, acceptor and non-splice site classes. We also performed comparative analysis with multiple baseline models. As expected, all models performed well on initial dataset as the task is oversimplified, although DNABERT-Splice still achieved the best (Supplementary Table S3). We, then, compared DNABERT-Splice with all baselines using a reconstructed dataset that includes 'adversarial examples' (Fig. 5a). This time, the predictive performance of the baseline models greatly dropped, while DNABERT-Splice still achieved best accuracy of 0.923, F1 of 0.919 and MCC of 0.871, with AUROC and AUPRC significantly greater than other models (Fig. 5b and c), which was also supported by Mcnemar's exact test (Supplementary Figs S10 and S11). Furthermore, DNABERT-Splice again outperformed all models when predicting on an independent test set containing held-out sliding-window scans from our iterative training process (Supplementary Table S4). We also examined the attention landscape to elucidate on how model made classification decision (Supplementary Fig. S12). Surprisingly, DNABERT-Splice showed globally consistent high attention upon intronic regions (downstream of donors and upstream of acceptors), highlighting the presence and functional importance of various intronic splicing enhancers (ISEs) and silencers (ISSs) acting as CREs for splicing (Wang and Burge, 2008).

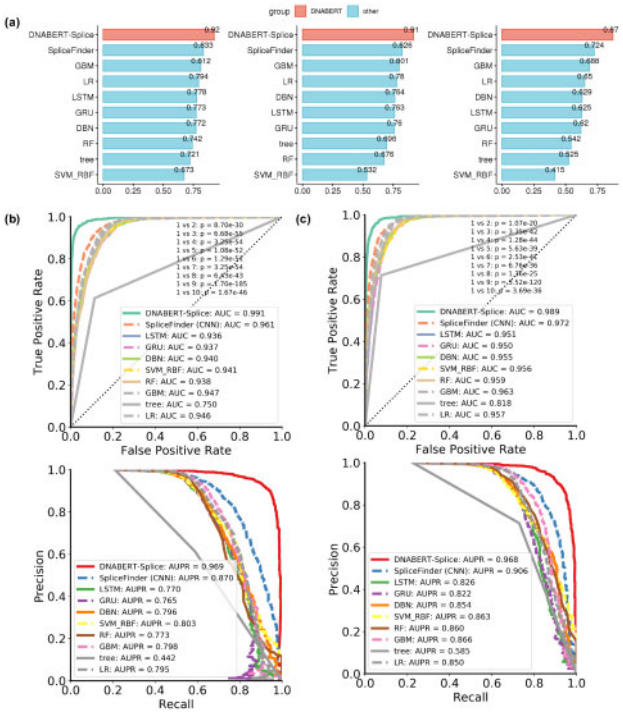

Fig. 5. DNABERT significantly outperforms other models in finding splice sites.(**a**) (Left to right) multiclass accuracy, F1 and MCC of splice donor and acceptor prediction. GBM: gradient boosting; LR: logistic regression; DBN: deep belief network; RF: random forest; tree: decision tree; SVM_RBF: support vector machine with radial basis function kernel. (**b**, **c**) ROC (top) and PR curves (bottom) on splice donor (b) and acceptor (c) datasets with adjusted *P*-values from Delong test

3.5 Identifying functional genetic variants with DNABERT

We applied DNABERT to identify functional variants using around 700 million short variants in dbSNP(Sherry, 2001). Specifically, we selected only those variants that are located inside DNABERT predicted high-attention regions and repeated the predictions, using sequences with altered alleles. Candidate variants resulting in significant changes in prediction probability were queried in ClinVar(Landrum *et al.*, 2014), GRASP (Leslie *et al.*, 2014) and NHGRI-EBI GWAS Catalog (Buniello *et al.*, 2019). In Prom-300 dataset, we found 24.7% and 31.4% of dbSNP Common variants we identified using TATA and non-TATA promoters are present in at least one of the three databases (Supplementary Table S6). We present some example functional variants that we found using ENCODE 690 ChIP-seq datasets (Fig. 6a–c). Figure 6a shows a rare, pathogenic 4 bp deletion completely disrupts a CTCF BS within MYO7A gene in ECC-1 cell line. This deletion is known to cause Usher Syndrome, an autosomal recessive disorder characterized by deafness, although the relationship with CTCF is to be determined (Jaijo *et al.*, 2006). Similarly, Figure 6b depicts how a rare single nucleotide variant (SNV) at initiator codon of SUMF1 gene, which leads to multiple sulfatase deficiency (Cosma *et al.*, 2003), simultaneously disrupts a YY1 BS with unknown functional consequences. In Figure 6c, a common risk variant of pancreatic cancer at intronic region of XPC gene also greatly weakens CTCF BS (Liang *et al.*, 2018). In all examples, DNABERT consistently shows highest attention at/around the variants of interest. We finally evaluated the quality of DNABERT-created mutational scores, comparing to those from other models, in globally prioritizing functional variants (Supplementary Methods). Using same set of functional SNVs from PRVCS benchmark dataset (Li *et al.*, 2016), model trained on mutation scores from DNABERT predictions on ENCODE 690 TF dataset achieves better AUROC than those using scores from other deep learning models (Supplementary Fig. S13). We expect the performance to be further enhanced as we bring in other features, such as DNase I hypersensitivity (DHS) predictions and others.

3.6 Pre-training substantially enhances performance and generalizes to other organisms

Lastly, we investigated the importance of pre-training based on performance enhancement and generalizability. When comparing training loss of pre-trained DNABERT-prom-300 with randomly initialized ones under same hyperparameters, pre-trained DNABERT converges to a markedly lower loss, suggesting that randomly initialized models get stuck at local minima very quickly without pre-training, as it ensures preliminary understanding of DNA logic by capturing distant contextual information (Fig. 6d). Similarly, randomly initialized DNABERT-prom-core models either remain completely untrainable or exhibit suboptimal performance. An examination of attention maps reveals the gradual comprehension of input sequence (Fig. 6e). Since separate pre-training of DNABERT for different organisms can be both very time-consuming and resource-intensive, we also evaluated whether DNABERT pre-trained on human genome can be also applied on other mammalian organisms. Specifically, we fine-tuned DNABERT pre-trained with human genome on 78 mouse ENCODE ChIP-seq datasets (Mouse *et al.*, 2012) and compared with CNN, CNN+LSTM, CNN+GRU and randomly initialized DNABERT. Pre-trained DNABERT significantly outperformed all baseline models (Fig. 6f), showing the robustness and applicability of DNABERT even across a different genome. It is well known that although the protein-coding regions between human and mouse genomes are approximately 85% orthologous, the non-coding regions only show approximately 50% global similarity (Mouse Genome Sequencing

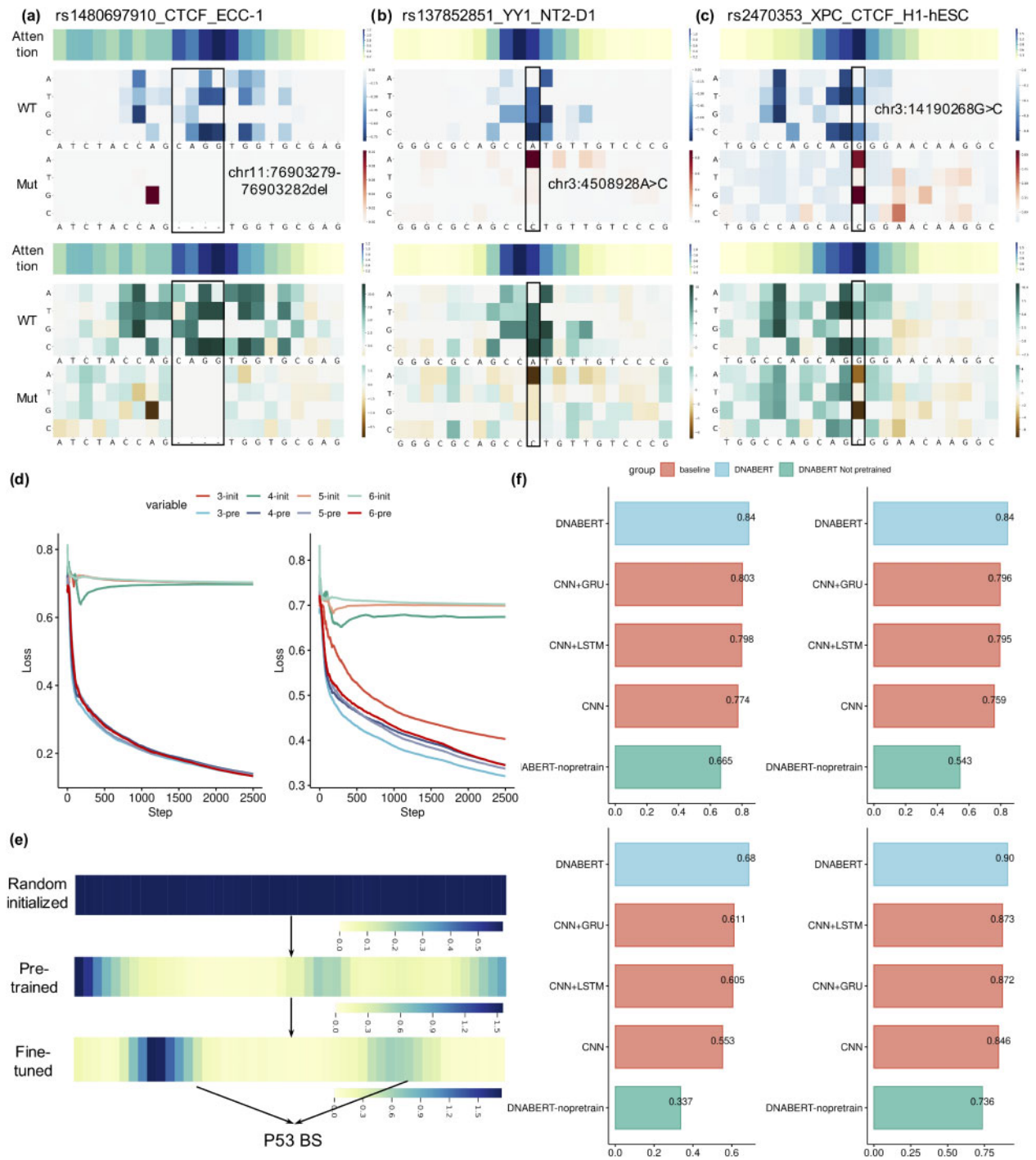

Fig. 6. DNABERT identifies functional genetic variants, and pretraining is essential and can be generalized.(**a–c**) Mutation maps of difference scores (top 3) and log-odds ratio scores (logOR, bottom 3). Each mutation map contains the attention score indicating importance of the region (top), scores for wild-type (WT, middle) and scores for mutant (mut, bottom). (Left to right) a rare deletion within a CTCF binding site inside MYO7A gene in ECC-1 cell line completely disrupts the binding site; a rare single-nucleotide variant (SNV) at initiator codon of SUMF1 gene also disrupts YY1 binding site (5'-CCGCCATNTT-3'); a common intronic SNP within XPC gene weakens CTCF binding site and is associated with pancreatic cancer. (**d**) Fine-tuning loss of pre-trained (pre) versus random initialized (init) DNABERT on Prom-300 (left) and Prom-core (right). (**e**) p53 attention map for random initialized (top), pre-trained (middle) and fine-tuned (bottom) DNABERT model. (**f**) Mean Accuracy (top left), F1 (top right), MCC (bottom left) and AUC (bottom right) across 78 mouse ENCODE datasets

et al., 2002). Since TFBS mostly locate within the non-coding region, DNABERT model successfully transferred learned information from one genome to a much less similar genome with very high tolerance to the dissimilarities. This demonstrates that the model correctly captured common deep semantics within DNA sequences across organisms. The evaluations above demonstrates the essentiality of pre-training and guarantees extensibility of the pre-trained model for efficient application in numerous sequence prediction tasks across different organisms.

4 Discussion

Transformers-based models have achieved state-of-the-art performance on various NLP tasks (Devlin *et al.*, 2018; Liu *et al.*, 2019;

Yang *et al.*, 2019)and for biomedical and clinical entity extraction from large-scale EHR notes (Li *et al.*, 2019) and biomedical documents (Lee *et al.*, 2020). Previous research has applied Transformers on protein sequences and prokaryotic genomes (Clauwaert and Waegeman, 2020; Min *et al.*, 2019). Here, we demonstrated that DNABERT achieved superior performance across various downstream DNA sequence prediction tasks by largely surpassing existing tools. Using an innovative global contextual embedding of input sequences, DNABERT tackles the problem of sequence specificity prediction with a 'top-down' approach by first developing general understanding of DNA language via self-supervised pre-training and then applying it to specific tasks, in contrast to the traditional 'bottom-up' approach using task-specific data. These characteristics of DNABERT ensures that it can more effectively learn from DNA context with great flexibility adapting to multiple situations, and enhanced performance with limited data. In particular, we also observed great generalizability of pre-trained DNABERT across organisms, which ensures the wide applicability of our method without the need for separate pre-training.

The pre-trained DNABERT model, released as part of this study, can be implemented for other sequence prediction tasks, for example, determining CREs and enhancer regions from ATAC-seq(Buenrostro *et al.*, 2013) and DAP-seq(Bartlett *et al.*, 2017). Further, since RNA sequences differs from DNA sequences only by one base (thymine to uracil), while the syntax and semantics remain largely the same, our proposed method can also apply to Cross-linking and immunoprecipitation (CLIP-seq) data for prediction of binding preferences of RNA-binding proteins (RBPs) (Gerstberger *et al.*, 2014). Although direct machine translation on DNA is not yet possible, the successful development of DNABERT shed light on this possibility. As a successful language model, DNABERT correctly captures the hidden syntax, grammar and semantics within DNA sequences and should perform equally well on Seq2seq translation tasks once token-level labels become available. Meanwhile, other aspects of resemblance between DNA and human language beyond text (e.g. alternative splicing and punctuation) highlights the need to combine data of different level for more proper deciphering of DNA language. In summary, we anticipate that DNABERT can bring new advancements and insights to the bioinformatics community by bringing advanced language modeling perspective to gene regulation analyses.

Acknowledgements

The authors thank Dr. Manoj Kandpal for his help with the initial processing of GEO datasets and reading of the manuscript.

Funding

This work was supported by National Library of Medicine/National Institutes of Health funding [R01LM011297 to R.D.].

Conflict of Interest: none declared.

References

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*,**33**, 831–838.

Andersson,R. and Sandelin,A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*,**21**, 71–87.

Bartlett,A. *et al.* (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.*,**12**, 1659–1672.

Bengio,Y. *et al.* (2013) Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal.*,**35**, 1798–1828.

Brendel,V. and Busse,H.G. (1984) Genome structure described by formal languages. *Nucleic Acids Res.*,**12**, 2561–2568.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*,**10**, 1213–1218.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*,**47**, D1005–D1012.

Cho,K. *et al.* (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In:* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734.

Clauwaert,J. and Waegeman,W. (2020) Novel transformer networks for improved sequence labeling in genomics. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Cosma,M.P. *et al.* (2003) The multiple sulfatase deficiency gene encodes an essential and limiting factor for the activity of sulfatases. *Cell*,**113**, 445–456.

Davuluri,R.V. (2003) Application of FirstEF to find promoters and first exons in the human genome. *Curr.Protoc.Bioinf.*,**29**, 412–417.

Davuluri,R.V. *et al.* (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.*,**24**, 167–177.

Devlin,J. *et al.* (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv : 1810.04805*.

Dreos,R. *et al.* (2013) EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.*,**41**, D157–D164.

Dunham,I. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*,**489**, 57–74.

Gerstberger,S. *et al.* (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*,**15**, 829–845.

Gibcus,J.H. and Dekker,J. (2012) The context of gene expression regulation. *F1000 Biol. Rep.*,**4**, 8.

Gupta,S. *et al.* (2007) Quantifying similarity between motifs. *Genome Biol.*,**8**, R24.

Hassanzadeh,H.R. and Wang,M.D. (2016) DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In: Proceedings of the *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 178–183.

Head,T. (1987) Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. *Bull. Math. Biol.*,**49**, 737–759.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*,**9**, 1735–1780.

Jaijo,T. *et al.* (2006) MYO7A mutation screening in Usher syndrome type I patients from diverse origins. *J. Med. Genet.*, **44**, e71.

Ji,S. (1999) The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Ann. N. Y. Acad. Sci. Paper Ed.*,**870**, 411–417.

Ji,Y. *et al.* (2020) In silico analysis of alternative splicing on drug–target gene interactions. *Sci. Rep.*,**10**, 134.

Kelley,D.R. *et al.* (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*,**26**, 990–999.

Khamis,A.M. *et al.* (2018) A novel method for improved accuracy of transcription factor binding site prediction. *Nucleic Acids Res.*,**46**, e72.

Koeppel,M. *et al.* (2011) Crosstalk between c-Jun and TAp73alpha/beta contributes to the apoptosis-survival balance. *Nucleic Acids Res.*,**39**, 6069–6085.

Landrum,M.J. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*,**42**, D980–D985.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*,**521**, 436–444.

Lee,J. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*,**36**, 1234–1240.

Leslie,R. *et al.* (2014) GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*,**30**, i185–i194.

Li,F. *et al.* (2019) Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. *JMIR Med. Inform.*,**7**, e14830.

Li,M.J. *et al.* (2016) Predicting regulatory variants with composite statistic. *Bioinformatics*,**32**, 2729–2736.

Liang,X.H. *et al.* (2018) Interaction of polymorphisms in xerodermapigmentosum group C with cigarette smoking and pancreatic cancer risk. *OncolLett.*,**16**, 5631–5638.

Liu,Y. *et al.* (2019) Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv : 1907.11692*.

Mantegna,R.N. *et al.* (1994) Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.*,**73**, 3169–3172.

Min,S. *et al.* (2019) Pre-training of deep bidirectional protein sequence representations with structural information. *arXiv preprint arXiv : 1912.05625*.

Mouse,E.C. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.*,**13**, 418.
Nirenberg,M. *et al.* (1965) RNA codewords and protein synthesis, VII. On the general nature of the RNA code.*Proc. Natl. Acad. Sci. USA*,**53**, 1161–1168.
Oubounyt,M. *et al.* (2019) DeePromoter: robust promoter predictor using deep learning. *Front. Genet.*,**10**, 286.
Quang,D. and Xie,X.H. (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*,**44**, e107.
Searls,D.B. (1992) The linguistics of DNA. *Am. Sci.*,**80**, 579–591.
Searls,D.B. (2002) The language of genes. *Nature*,**420**, 211–217.
Shen,Z. *et al.* (2018) Recurrent neural network for predicting transcription factor binding sites. *Sci. Rep. UK*,**8**, 1–10.
Sherry,S.T. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*,**29**, 308–311.
Solovyev,V. *et al.* (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.*,**7**, S10–12.
Umarov,R. *et al.* (2019) Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics*,**35**, 2730–2737.
Vaswani,A. *et al.* (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010.
Vitting-Seerup,K. and Sandelin,A. (2017) The landscape of isoform switches in human cancers. *Mol. Cancer Res.*,**15**, 1206–1220.
Wang,R.H. *et al.* (2019) SpliceFinder: ab initio prediction of splice sites using convolutional neural network. *BMC Bioinformatics*,**20**, 652.
Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*,**14**, 802–813.
Waterston,R.H. *et al.*; Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*,**420**, 520–562.
Yang,Z. *et al.* (2019) Xlnet: generalized autoregressive pretraining for language understanding. In: *Proceedings of the 32nd International Conference* on Neural Information Processing Systems, pp. 5754–5764.
Yoon,H. *et al.* (2002) Gene expression profiling of isogenic cells with different TP53 gene dosage reveals numerous genes that are affected by TP53 dosage and identifies CSPG2 as a direct target of p53. *Proc. Natl. Acad. Sci. USA*,**99**, 15632–15637.
Zhang,Y.Q. *et al.* (2020) DeepSite: bidirectional LSTM and CNN models for predicting DNA-protein binding. *Int. J. Mach. Learn. Cyb.*,**11**, 841–851.
Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*,**12**, 931–934.
Zou,J. *et al.* (2019) A primer on deep learning in genomics. *Nat. Genet.*,**51**, 12–18.