Prediction of RNA-protein interactions using a nucleotide language model

Keisuke Yamada[1] and Michiaki Hamada[1,2*]

[1]School of Advanced Science and Engineering, Waseda University, Tokyo, Japan
[2]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
[*]To whom correspondence should be addressed.

Abstract

Motivation: The accumulation of sequencing data has enabled researchers to predict the interactions between RNA sequences and RNA-binding proteins (RBPs) using novel machine learning techniques. However, existing models are often difficult to interpret and require additional information to sequences. Bidirectional encoder representations from Transformer (BERT) is a language-based deep learning model that is highly interpretable. Therefore, a model based on BERT architecture can potentially overcome such limitations.
Results: Here, we propose BERT-RBP as a model to predict RNA-RBP interactions by adapting the BERT architecture pre-trained on a human reference genome. Our model outperformed state-of-the-art prediction models using the eCLIP-seq data of 154 RBPs. The detailed analysis further revealed that BERT-RBP could recognize both the transcript region type and RNA secondary structure only from sequential information. Overall, the results provide insights into the fine-tuning mechanism of BERT in biological contexts and provide evidence of the applicability of the model to other RNA-related problems.
Availability: Python source codes are freely available at https://github.com/kkyamada/bert-rbp.
Contact: mhamada@waseda.jp

1 Introduction

Interactions between RNA sequences and RNA-binding proteins (RBPs) have a wide variety of roles in regulating cellular functions, including mRNA modification, splicing, translation, and localization (Hentze *et al.*, 2018). For instance, the T-cell-restricted intracellular antigen family of proteins function as alternative splicing regulators (Wang *et al.*, 2010), and heterogeneous nuclear ribonucleoprotein K (hnRNPK) is a versatile regulator of RNA metabolism (Geuens *et al.*, 2016). Numerous attempts have been made to identify RNA-RBP interactions to accurately capture their biological roles.

Among the various *in vivo* experimental methods, high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (CLIP-seq) is widely used to reveal a comprehensive picture of RNA-RBP interactions (Licatalosi *et al.*, 2008; Lin and Miles, 2019). Altered CLIP-seq protocols have also been developed (Hafner *et al.*, 2010; König *et al.*, 2010; Van Nostrand *et al.*, 2016). Recently, a large amount of enhanced CLIP-seq (eCLIP) data, targeting more than 150 different RBPs, was generated during phase III of the Encyclopedia of DNA Elements (ENCODE) Project (Van Nostrand *et al.*, 2020a).

Because there is a vast volume of available CLIP-seq data, recent bioinformatics studies have focused on developing machine learning models to predict RNA-RBP interactions and deciphering hidden patterns translated by these models (Pan *et al.*, 2019; Yan and Zhu, 2020). Early models use statistical evaluations or support vector machines (SVMs) to classify RNA sequences into RBP-bound or RBP-unbound groups (Hiller *et al.*, 2006; Kazan *et al.*, 2010; Maticzka *et al.*, 2014). One of the SVM-based models, GraphProt, encodes RNA sequences and their estimated secondary structures into graph representations (Maticzka *et al.*, 2014). Non-negative matrix factorization (NMF) and random forest were also adapted to other models (Stražar *et al.*, 2016; Yu *et al.*, 2019). Since Alipanahi *et al.* (2015) demonstrated the applicability of convolutional neural networks (CNNs) for predicting RNA-protein and DNA-protein interactions, several deep learning models have been developed. While some models incorporate a single CNN with some modifications (Pan and Shen, 2018; Zhang *et al.*, 2019; Tahir *et al.*, 2021), others use a different neural network model (Uhl *et al.*, 2020) or a combination of several neural network architectures (Ben-Bassat *et al.*, 2018; Pan *et al.*, 2018; Yan *et al.*, 2020; Deng *et al.*, 2020; Grønning *et al.*, 2020). For instance, HOCNNLB uses high-order encodings of RNA sequences as inputs for CNN (Zhang *et al.*, 2019), and iDeepS uses stacked CNN and bidirectional long short-term memory (biLSTM) and takes both RNA sequences and their estimated secondary structures as inputs (Pan *et al.*, 2018). However, existing models are often poorly interpretable because of the complex nature of the neural network and require additional information to RNA sequences. Therefore, the development of advanced models that overcome these limitations is awaited.

The improvement of deep learning architectures largely buttresses progress in building better bioinformatics tools. In the field of natural language processing, self-attention-based deep learning architectures, such as Transformer and BERT, have achieved state-of-the-art performance in vari-

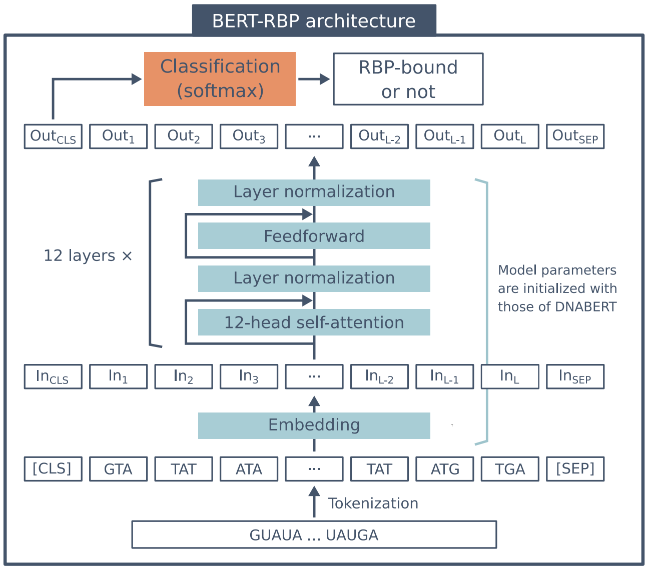

Figure 1: The architecture of BERT-RBP. The input RNA sequence was first tokenized into 3-mers and modified with CLS (classification) and SEP (separation) tokens. Then, each token was embedded into a 768-dimensional feature vector. These feature vectors were consequently processed through 12 Transformer encoder layers, where each layer included 12 self-attention heads. The CLS token of the output vector from the last layer was further utilized for classification to predict whether the input RNA sequence bound to the RBP. Upon fine-tuning, the model parameters of the embedding layer and the stacked Transformer encoder layers were initialized with those of DNABERT. These parameters were randomly initialized for the BERT-baseline model.

ous tasks (Vaswani *et al.*, 2017; Devlin *et al.*, 2018). Additionally, BERT, which essentially consists of stacked Transformer encoder layers, shows enhanced performance in downstream task-specific predictions after pre-training on a massive dataset (Devlin *et al.*, 2018). In the field of bioinformatics, several BERT architectures pre-trained on a massive corpus of protein sequences have been recently proposed, demonstrating their capability to decode the context of biological sequences (Rao *et al.*, 2019; Rives *et al.*, 2020; Elnaggar *et al.*, 2020; Iuchi *et al.*, 2021). In comparison to the protein language models, Ji *et al.* (2021) a pre-trained BERT model, named DNABERT, on a whole human reference genome demonstrated its broad applicability for predicting promoter regions, splicing sites, and transcription factor binding sites upon fine-tuning. Thus, pre-trained BERT models are potentially advantageous for a wide variety of bioinformatics tasks, including the prediction of RNA-RBP interactions.

In addition to its performance, BERT is highly interpretable and suitable for translating extended contextual information compared to conventional deep learning architectures, such as CNNs and long short-term memory (Rogers *et al.*, 2020). Researchers in an emerging field, called BERTology, intend to elucidate how BERT learns contextual information by analyzing attention, which essentially represents the flow of information within a model (Vig and Belinkov, 2019). For instance, analysis of protein BERT models revealed that protein contact maps could be reconstructed from the attention of the model (Vig *et al.*, 2021; Rao *et al.*, 2021). This implies that, by analyzing the fine-tuned BERT model, we can reasonably explain the types of features that are crucial for predicting RNA-RBP interactions.

In this study, we applied the BERT model pre-trained on a human reference genome to predict the RBP-binding property of RNA sequences. Our model, named BERT-RBP, outperformed existing state-of-the-art models as well as the baseline BERT model whose weight parameters were randomly initialized, showing the significance of pre-training on a large corpus. Attention analysis on the fine-tuned model further revealed that BERT-RBP could translate biological contexts, such as transcript region type, transcript region boundary, and RNA secondary structure, only from RNA sequences. Thus, this study highlights the powerful capability of BERT in predicting RNA-RBP interactions and provides evidence of the architecture's potential applicability to other bioinformatics problems.

2 Materials and methods

2.1 Terminology

k-mer : For a given sequence, k-mers of the sequence consisted of every possible subsequence with length k, i.e., given a sequence $ACGTAC$, the 3-mers of this sequence included ACG, CGT, GTA, and TAC, and the 4-mers included $ACGT$, $CGTA$, and $GTAC$.

Token : Tokens are referred to as words or their positions within a sequence. Tokens included not only k-mers but also special tokens, such as CLS (classification) and SEP (separation). In our model, CLS was appended at the beginning of each input sequence, and its feature vector from the final layer was used for classification. The SEP was attached only to the end of each sequence.

Attention : Attention represents the flow of information within the BERT model. The attention weight indicates how much information the hidden state of a token in the upper (closer to the output) layer referred to the hidden state of a token in the lower (closer to the input) layer.

2.2 Data Preparation

An eCLIP-seq dataset previously generated from the ENCODE3 database by Pan *et al.* (2020) was used. The original dataset consisted of 154 RBP sets with up to 60,000 positive RNA sequences that bind to the corresponding RBP and the same number of negative sequences. Each positive sequence had a length of 101 nucleotides with the eCLIP-seq read peak at its center, while each negative sequence was sampled from the non-peak region of the same reference transcript as its positive counterpart. First, all sequences that included unannotated regions or repeatedly appeared in the same set were removed to create the dataset, and 12,600 positive and negative sequences were randomly sampled. If the original RBP set included less than 12,600 samples, all sequences were retrieved. The resulting samples were split into training (19,200) and test sets (6,000). Additionally, for the selected RBPs, non-training datasets were created that included all positive and negative samples, except those in the training sets. The training sets were used during the fine-tuning step, the individual test sets were used to measure performance after fine-tuning, and the non-training sets were used for attention analysis.

2.3 Models and Training

2.3.1 Pre-trained BERT model

DNABERT, a BERT-based architecture pre-trained on a human reference genome, was adapted to model RNA sequences and their RBP-binding properties. Briefly, DNABERT was pre-trained on k-mer (k = 3-6) representations of nucleotide sequences obtained from a human reference genome, GRCh38.p13 (Ji *et al.*, 2021). Once the CLS and SEP tokens were appended to the input k-mers, each token was embedded into real vectors with 768 dimensions. The model was pre-trained with the masked language modeling objective, self-supervised learning, to predict randomly masked tokens using information from other tokens. The model had 12 Transformer encoder layers, each of which consisted of 12 self-attention heads and utilized the multi-head self-attention mechanism (Figure 1).

2.3.2 Fine-tuning

Upon fine-tuning, the parameters of the model were initialized with those of DNABERT (Figure 1). Subsequently, BERT-RBP was fine-tuned on the training datasets. The hyperparameters used for training are listed in Table S1, and these hyperparameters were kept consistent for all the different k-mer models (k = 3-6). The models were trained on four NVIDIA Tesla V100 GPUs (128GB memory). The training of one RBP model using 19,200 samples took less than 10 min. After fine-tuning, the model performance was measured using the area under the receiver operating characteristic curve (AUROC) using independent test sets.

2.3.3 Baseline Models

The following three existing models were implemented as baselines: GraphProt, iDeepS, and HOCNNLB. GraphProt is an SVM-based model that converts RNA sequences and their estimated secondary structures into graph representations and predicts RBP-binding sites (Maticzka *et al.*, 2014). iDeepS uses a combination of CNN and biLSTM to predict RBP-binding sites from RNA sequences and their estimated secondary structures (Pan *et al.*, 2018). HOCNNLB is another method for training CNNs to predict RBP binding sites while taking k-mer representations of RNA sequences (Zhang *et al.*, 2019). In addition to the above models, the baseline BERT model (BERT-baseline), whose parameters were randomly initialized instead of transferring parameters from DNABERT, was also trained. Hyperparameters were kept consistent with BERT-RBP except that the learning rate was set to a ten times larger value (0.002) to promote optimization. All baseline models were trained and tested using the same training and independent test sets as BERT-RBP.

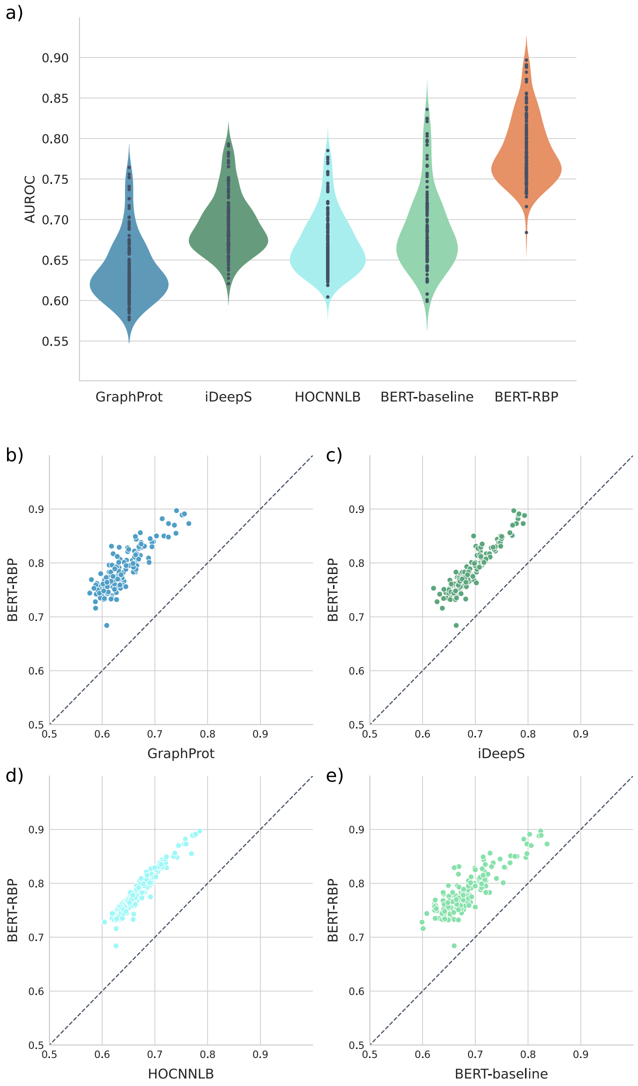

Figure 2: Overall performance of our model. a) Area under the receiver operating characteristic curve (AUROC) scores of BERT-RBP and four baseline models over 154 RBP datasets. Each violin plot shows the performance of each model, and each dot within each violin plot represents the AUROC score for a single RBP dataset. b-e) Detailed comparison of this model's performance against b) GraphProt, c) iDeepS, d) HOCNNLB, or e) BERT-baseline by AUROC measurement. Each dot represents the AUROC scores of BERT-RBP and the corresponding baseline model trained using the same RBP dataset. The diagonal dashed line indicates that the performances of the two models are identical.

2.4 Attention Analysis

We examined whether attention reflected any biological features of the input RNA sequences after fine-tuning. The method proposed by Vig *et al.* (2021) was adapted to ask whether attention agrees with hidden properties of inputs both at the sequence level (transcript region type) and at the token level (transcript region boundary and RNA secondary structure).

2.4.1 Sequence-level and Token-level Properties

When conditioned by an input sequence, each head emits a set of attention weights α, where $\alpha_{i,j} (> 0)$ indicates the attention weight from the ith token in the upper layer to the jth token in the lower layer. $\sum_j \alpha_{i,j} = 1$ is satisfied, as the attention weights are normalized over each token in the upper layer. We calculated the attention weights for the CLS in each head as follows:

$$s_\alpha(f) = \frac{\frac{1}{M}\sum_{m=1}^{M} f(m) \sum_{i}^{L+2} \alpha_{i,\mathrm{CLS}}}{\frac{1}{N-M}\sum_{m=1}^{N-M}(1-f(m))\sum_{i}^{L+2}\alpha_{i,\mathrm{CLS}}} \quad (1)$$

where N and M indicate the number of input sequences and the number of inputs with the property of interest, respectively; $f(m)$ is an indicator that returns 1 if the property exists in the mth sequence and 0 otherwise; L indicates the number of sequential tokens in the mth sequence, and $L+2$ is the number of all tokens, including CLS and SEP. Intuitively, $s_\alpha(f)$ represents the relative attention to the CLS associated with the property f.

For token-level analysis, attention weights to the token of interest were computed at each head using the following equation:

$$t_\alpha(g) = \frac{\sum_{m=1}^{M}\sum_{i}^{L}\sum_{j}^{L} g(j)\alpha_{i,j}}{\sum_{m=1}^{N}\sum_{i}^{L}\sum_{j}^{L}\alpha_{i,j}} \quad (2)$$

where $g(j)$ is an indicator that returns 1 if the property exists in the jth token in the lower layer and 0 otherwise. Note that attention weights to CLS and SEP were not considered during the token-level analysis, and the sequential token length L was used. Here, $t_\alpha(g)$ represents the ratio of attention to property g.

2.4.2 Analysis of Transcript Region Type

We first examined whether attention weights reflect transcript region types, including the 5'UTR, 3'UTR, intron, and CDS. Region-type annotations were downloaded from the Ensembl database (Ensembl Genes 103, GRCh38.p13) (Yates *et al.*, 2020). For each gene, we selected the most prominent isoform based on the APPRIS annotation (Rodriguez *et al.*, 2013), transcript support level, and length of the transcript (the longer, the better). Region types were applied for each nucleotide as binary labels, resulting in a 4×101 annotation matrix per sequence. The original eCLIP dataset curated by (Pan *et al.*, 2020) used GRCh37/hg19 as a reference genome, so we converted sequence positions into those of GRCh38/hg38 using the UCSC liftOver tool (Kent *et al.*, 2002) and retained those sequences that could be remapped with 100% sequence identity. For simplicity, sequences containing one or more nucleotides labeled with the region type were regarded as having that property. Using the non-training dataset, we accumulated attention weights to the CLS token at each head, averaged over the region type, and calculated the attention level relative to the background (Equation (1)). Consequently, the head, which showed the most significant relative attention level, was selected for each RBP and each region type, and the raw attention weights to CLS ($\sum_i \alpha_{i,\mathrm{CLS}}$) were extracted from the head of each sample. Because of the nature of attention, the number of samples tends to be sparse in the range where attention weights are relatively high; therefore, samples whose attention weights were within the 99.5 percentile were the focus. Finally, the Spearman's rank correlation coefficient between the raw attention weights to CLS and the RNA sequence probability of being the region type were calculated.

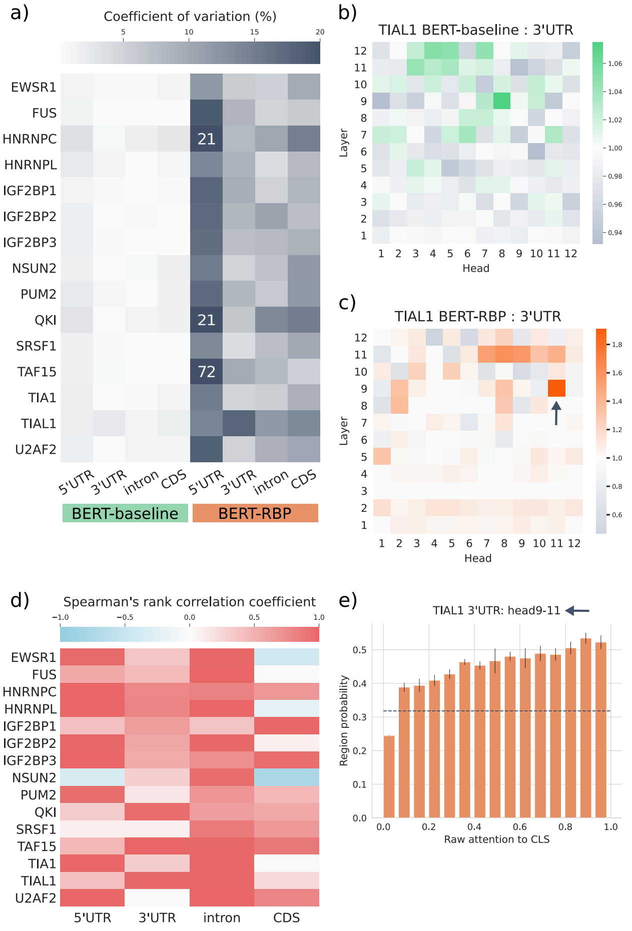

Figure 3: Results of sequence-level attention analysis of transcript region type. a) The degree of specialization was measured for 15 RNA-binding proteins (RBPs) and four region types using BERT-baseline and BERT-RBP. The degree of specialization was evaluated using the coefficient of variation of the relative attention levels among 144 attention heads. The value was directly annotated for data points where the coefficient of variation was saturated (¿20%).b-c) Exemplary results showing attention patterns measured by the relative attention to CLS among 144 heads. b) BERT-baseline and c) BERT-RBP trained on the same TIAL1 training set were analyzed using the 3'UTR annotation. d) Correlation analysis was conducted between the raw attention weights to CLS and the RNA sequence probability of the region type. For each RBP and region type, the head showing the greatest relative attention (most specialized) was chosen, and the Spearman's rank correlation coefficient was calculated. e) An example showing the relationship between raw attention to CLS and the RNA sequence probability as the region type. For the BERT-RBP trained on the TIAL1 training set, the head 9-11 was selected to be most specialized for the 3'UTR in the heatmap (specified with the arrow in Figure 3c). The horizontal dashed line represents the background probability of the 3'UTR label within the TIAL1 non-training dataset. Error bars represent the mean ± standard deviations among three subsets randomly split from the original non-training set.

2.4.3 Analysis of Transcript Region Boundary

In addition, analysis of the transcript region boundary was conducted using non-training datasets with region-type annotations. 3-mer tokens containing nucleotides labeled with different region types were defined as having boundaries. Instead of the relative attention level, the ratio of attention weights to the property (Equation (2)) was calculated for each head so that the calculated ratio could be compared to the overall probability of the region boundary within the dataset. The similarity of attention ratio patterns between DNABERT and BERT-RBP was measured using Pearson's correlation coefficient to assess the degree of variation from DNABERT to BERT-RBP. Finally, the correlation between the raw attention weight to the boundary and the 3-mer token probability of the boundary was analyzed using the same pipeline used for transcript region type analysis.

2.4.4 Analysis of RNA Secondary Structure

The secondary structure of RNA was another property that was analyzed. For each input RNA sequence, the structure was estimated based on the maximum expected accuracy (MEA) using LinearPartition (Zhang *et al.*, 2020). Once the MEA structures were estimated, each nucleotide was labeled with one of six structural properties; F (dangling start), T (dangling end), I (internal loop), H (hairpin loop), M (multi-branched loop), and S (stem). 3-mer tokens containing one or more nucleotides labeled with structural properties were defined as having the structure. Similar to the transcript region boundary analysis, the ratio of attention weights to the structural property (Equation (2)) was computed for each head and compared to the overall probability of the structure within the dataset. Consequently, we analyzed the similarity of attention patterns between DNABERT and BERT-RBP and the correlation between the raw attention weight to the structure and the 3-mer token probability to have the structural property.

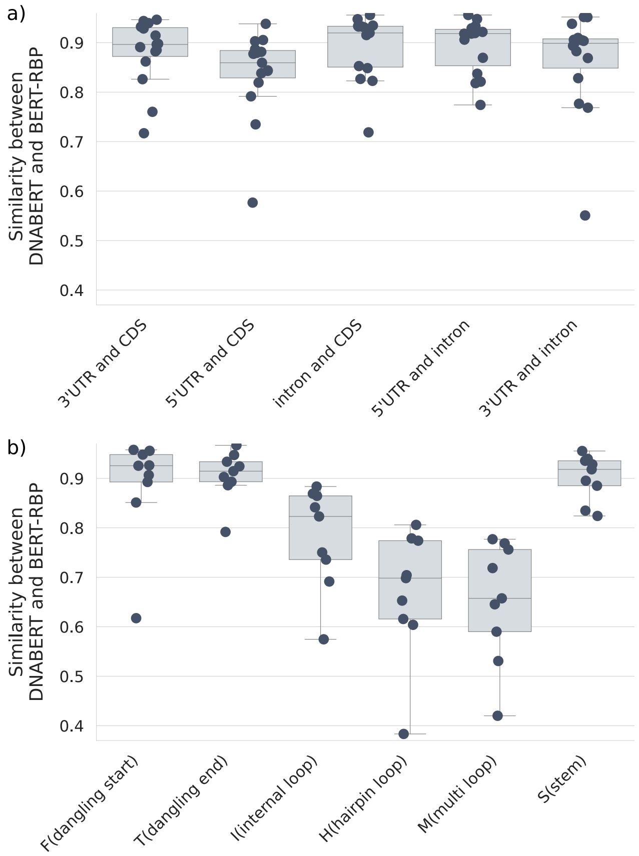

Figure 4: Results of token-level attention analysis of a) the transcript region boundary and b) the RNA secondary structure. The similarity between DNABERT and BERT-RBP was measured for a) five transcript region boundaries and b) six structure types. The similarity was evaluated using Pearson's correlation coefficient between the attention ratio patterns of DNABERT and BERT-RBP. Each dot represents the similarity score for a single RBP dataset.

3 Results

3.1 Performance of BERT-RBP

We evaluated the prediction performance of our model along with three existing models (GraphProt, iDeepS, and HOCNNLB) and BERT-baseline. GraphProt is an SVM-based model for predicting RNA-RBP interactions using graph representations of RNA sequences and their estimated secondary structures. iDeepS is based on stacked CNN and biLSTM and takes the sequence and estimated secondary structure of input RNAs to estimate their RBP binding properties. HOCNNLB uses k-mer representations of RNA sequences and predicts RNA-RBP interactions using a set of CNNs. All models were trained on the same training set, and their performance was measured using an independent test set over 154 RBPs. BERT-RBP resulted in an average AUROC of 0.785, which was higher than that of any other model (Figure 2a). The average AUROCs of the other models were 0.638, 0.693, 0.671, and 0.686 for GraphProt, iDeepS, HOCNNLB, and BERT-baseline, respectively. While the baseline BERT model alone showed comparable performance to existing methods, our model improved the scores, indicating the significance of pre-training on a large DNA corpus to predict RNA-RBP interactions. In addition, the score of BERT-RBP was higher than the previously reported AUROC (0.781) of the updated iDeepS (Pan *et al.*, 2020), even though we used approximately five times smaller subsets of their training data to train BERT-RBP. Furthermore, a one-to-one comparison against each baseline revealed that our model improved the scores for every single RBP dataset by a notable margin (Figure 2b-e). These results demonstrated that our model exceeded the state-of-the-art methods in predicting RNA-RBP interactions while taking only sequential information.

Because the original DNABERTs were pre-trained on 3- to 6-mer representations, we fine-tuned three other models, where each model takes 4- to 6-mer representations as inputs. When the AUROCs of fine-tuned models with different k-mers were compared, all fine-tuned models showed comparable performance to the 3-mer model, again demonstrating the robustness of the two-step training method (Figure S1). The detailed comparison showed that the 3-mer model outperformed others for 135 out of 154 RBPs; thus, we refer to the 3-mer model as BERT-RBP throughout this study.

3.2 Attention analysis

While being a deep learning model, BERT has high interpretability (Rogers *et al.*, 2020). In this study, we investigated

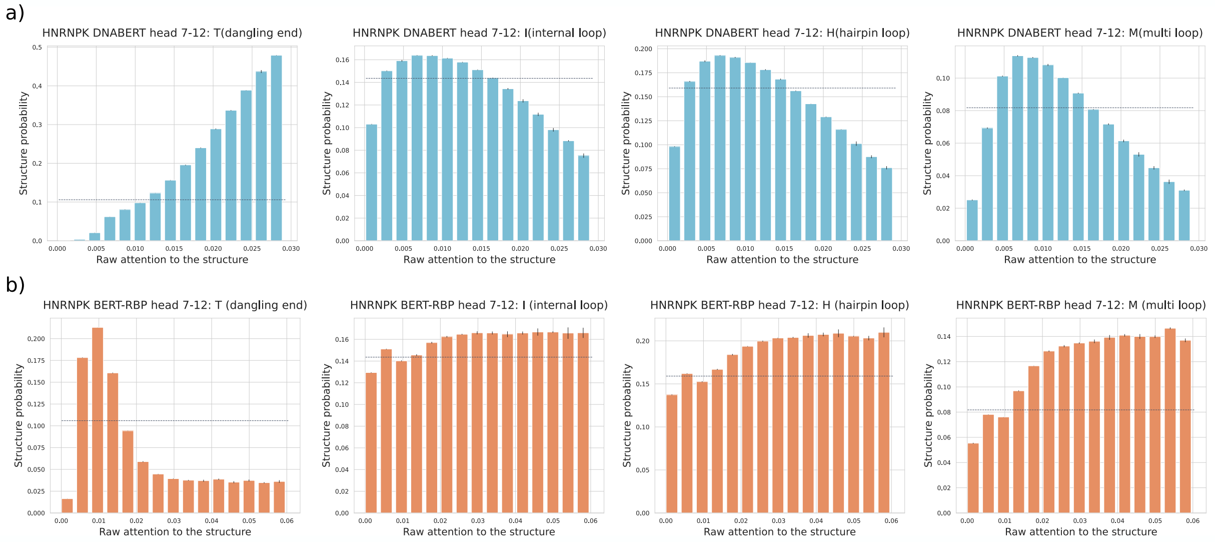

Figure 5: The detailed attention analysis of RNA secondary structure. a-b) Exemplary results of analysis for a) DNABERT and b) BERT-RBP using the same heterogeneous nuclear ribonucleoprotein K (hnRNPK) dataset. The relationship between the raw attention and the token probability for each structural property was measured at the head 6-8. The head was most specialized for detecting the internal, hairpin, and multi-branched loops within the BERT-RBP trained on the hnRNPK training set. The horizontal dashed lines represent the background probability of the corresponding structure within the hnRNPK non-training dataset. Error bars represent means ± standard deviations among three subsets randomly split from the original non-training set.

the types of biological information that could be deciphered by our model.

3.2.1 Transcript Region Type

The transcript region type plays an essential role in predicting RNA-RBP interactions (Stražar *et al.*, 2016; Avsec *et al.*, 2018; Uhl *et al.*, 2020). We investigated whether the model showed high attention toward sequences from a specific transcript region, such as 5'UTR, 3'UTR, intron, or CDS, as described in section 2.4.2. The use of the CLS token stimulates the model to accumulate sequence-level information in the CLS token (Rogers *et al.*, 2020). We hypothesized that attention associated with the CLS token represents sequence-level information, that is, the transcript region type property. To test this hypothesis, we computed the relative attention to CLS associated with each region type for both BERT-RBP and BERT-baseline using 15 selected RBP datasets. These 15 RBPs were selected because their CLIP-seq data were analyzed and included in the previous benchmark created by Stražar *et al.* (2016). When the degree of specialization was then measured using the coefficient of variation, the relative attention level of BERT-RBP varied more than that of BERT-baseline for all 15 RBPs and four transcript region types (Figure 3a-c and S2). This result indicated that BERT-RBP recognized the transcript region type of RNA sequences more than BERT-baseline.

To provide more evidence that our model could deduce transcript region types from RNA sequences, we inspected BERT-RBP heads, which showed the most substantial relative attention level, and calculated the probability of RNA sequences incorporating the corresponding transcript region type. Positive correlations were observed between the raw attention to CLS and the RNA sequence probability as the region type (Figure 3d-e). A strong correlation (¿ 0.9) for introns was observed in nine out of 15 RBPs and a positive correlation (¿ 0.6) for 14 RBPs. These correlations are attributable to the abundance of introns within the genome, as well as the eCLIP-seq datasets. QKI, TAF15, and TIAL1 exhibited strong correlations (¿ 0.9) for 3'UTR. Direct eCLIP-seq analysis and other experimental results provided evidence of the enrichment of binding sites in the 3'UTR and its functional importance for QKI, TAF15, and TIAL1 (Kapeli *et al.*, 2016; Meyer *et al.*, 2018; Ciolli Mattioli *et al.*, 2019; Van Nostrand *et al.*, 2020a,b). In addition, the correlations of TAF15 and TIAL1 agreed with the essential region types to predict RNA-RBP interactions previously reported using the NMF-based model (Stražar *et al.*, 2016). The correlation for CDS was above 0.9 in IGF2BP1 and IGF2BP3. These results are also in line with the known function of IGF2BPs of binding to the end of CDS and 3'UTR to stabilize messenger RNAs (Huang *et al.*, 2018). On the other hand, correlations for the 5'UTR were not in accordance with those of previous studies. This may explain the information translated by BERT-RBP but not by other models; however, more extensive rationalization is necessary.

3.2.2 Transcript Region Boundary

To further investigate the capability of the model to extract transcript region type information from sequences, we analyzed the boundary between transcript regions (Section 2.4.3). For most RBPs and region boundaries, the similarity of attention patterns between DNABERT and BERT-RBP was consistently above 0.7, indicating that the variation from DNABERT to BERT-RBP was limited during fine-tuning (Figure 4a). A detailed comparison of the most specialized head demonstrated that both models were analogously specialized for the transcript region boundary, especially among intron-exon boundaries (Figure S3). These results indicated that BERT-RBP's capability to translate transcript region boundaries was transferred from DNABERT, inferring the significance of pre-training using the genome-size corpus and the potential applicability of DNABERT for other prediction tasks where transcript region type information is crucial.

3.2.3 RNA Secondary Structure

The RNA secondary structure is another feature that improves the prediction performance of several models (Maticzka *et al.*, 2014; Stražar *et al.*, 2016; Chung and Kim, 2019; Deng *et al.*, 2020). Accordingly, we investigated whether our model could consider the RNA secondary structure during prediction (Section 2.4.4). For this purpose, nine RBPs with varied structural preferences were selected (Dominguez *et al.*, 2018; Adinolfi *et al.*, 2019). When DNABERT and BERT-RBP were compared, there were variations in the attention patterns for loop structures (Figure 4b). Using the hnRNPK dataset, we further examined the variation of attention weights at the head 7-12, which had the highest attention ratio for the internal, hairpin, and multi-branched loops. The examination revealed a shift in specialization from DNABERT to BERT-RBP (Figure 5a-b). The head 7-12 of DNABERT was initially specialized for detecting the dangling ends, but it began to attend more to loop structures after fine-tuning. The shifted specialization was also observed for the other seven RBPs (Figure S4). These results align with the RBP's general binding preferences toward unstructured regions (Dominguez *et al.*, 2018). Taken together, the pre-trained BERT architecture can vary the type of structural information processed during fine-tuning.

4 Discussion

Although our analysis implied that the fine-tuned model could utilize the information learned by DNABERT, it is necessary to conduct a more extensive BERTological analysis to elucidate the accurate picture of the fine-tuning mechanism of biological BERT models. The syntactic relationship among tokens, for example, is an intensely researched topic in BERTology and may incorporate hidden contextual patterns of nucleotide sequences (Goldberg, 2019). In the context of protein BERT models, it was recently demonstrated that protein contact maps can be reconstructed using the attention maps extracted from the pre-trained protein BERT model (Rao *et al.*, 2021). If one could overcome the difference in the frequency of tokens in contact, it would be possible to reconstruct the base-pairing probability matrix using the attention maps of a nucleotide BERT model.

In this study, we proposed BERT-RBP, a fine-tuned BERT model for predicting RNA-RBP interactions. Using the eCLIP-seq data of 154 different RBPs, our model outperformed state-of-the-art methods and the baseline BERT model. Attention analysis revealed that BERT-RBP could distinguish both the transcript region type and RNA secondary structure using only sequential information as inputs. The results also inferred that the attention heads of BERT-RBP could either utilize information acquired during DNABERT pre-training or vary the type of information processed when necessary. As the analysis demonstrated the model's capability to translate transcript region type and RNA secondary structure, DNABERT can potentially be applied to other RNA-related tasks, such as RNA subcellular localization prediction (Gudenas and Wang, 2018; Yan *et al.*, 2019), RNA secondary structure prediction (Chen *et al.*, 2020; Sato *et al.*, 2021), and RNA coding potential prediction (Hill *et al.*, 2018). Thus, this study provides a state-of-the-art tool to predict RNA-RBP interactions and infers that the same method can be applied to other bioinformatics tasks.

Acknowledgements

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. We acknowledge Dr. Shitao Zhao and Taro Matsutani for their support and comments on the manuscript.

Funding

This work was supported by the Ministry of Education, Culture, Sports, Science, and Technology (KAKENHI) [grant number: JP17K20032, JP16H05879, JP16H06279, and JP20H00624 to MH] and JST CREST [grant numbers: JPMJCR1881 and JPMJCR21F1 to MH].

References

Adinolfi, M. *et al.* (2019). Discovering sequence and structure landscapes in RNA interaction motifs. *Nucleic acids research*, **47**(10), 4958–4969.

Alipanahi, B. *et al.* (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology*, **33**(8), 831–838.

Avsec, Ž. *et al.* (2018). Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. *Bioinformatics*, **34**(8), 1261–1269.

Ben-Bassat, I. *et al.* (2018). A deep neural network approach for learning intrinsic protein-RNA binding preferences. *Bioinformatics*, **34**(17), i638–i646.

Chen, X. *et al.* (2020). RNA secondary structure prediction by learning unrolled algorithms. In *International Conference on Learning Representations*.

Chung, T. and Kim, D. (2019). Prediction of binding property of RNA-binding proteins using multi-sized filters and multi-modal deep convolutional neural network. *PloS one*, **14**(4), e0216257.

Ciolli Mattioli, C. *et al.* (2019). Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. *Nucleic acids research*, **47**(5), 2560–2573.

Deng, L. *et al.* (2020). Deep neural networks for inferring binding sites of RNA-binding proteins by using distributed representations of RNA primary sequence and secondary structure. *BMC genomics*, **21**(13), 866.

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

Dominguez, D. *et al.* (2018). Sequence, structure, and context preferences of human RNA binding proteins. *Molecular cell*, **70**(5), 854–867.e9.

Elnaggar, A. *et al.* (2020). ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, page 2020.07.12.199554.

Geuens, T. *et al.* (2016). The hnRNP family: insights into their role in health and disease. *Human genetics*, **135**(8), 851–867.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. *arXiv*, page 1901.05287.

Grønning, A. G. B. *et al.* (2020). DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning. *Nucleic acids research*, **48**(13), 7099–7118.

Gudenas, B. L. and Wang, L. (2018). Prediction of LncRNA subcellular localization with deep learning from sequence features. *Scientific reports*, **8**(1), 16385.

Hafner, M. *et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**(1), 129–141.

Hentze, M. W. *et al.* (2018). A brave new world of RNA-binding proteins. *Nature reviews. Molecular cell biology*, **19**(5), 327–341.

Hill, S. T. *et al.* (2018). A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic acids research*, **46**(16), 8105–8113.

Hiller, M. *et al.* (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research*, **34**(17), e117.

Huang, H. *et al.* (2018). Recognition of RNA n6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nature cell biology*, **20**(3), 285–295.

Iuchi, H. *et al.* (2021). Representation learning applications in biological sequence analysis. *bioRxiv*, page 2021.02.26.433129.

Ji, Y. *et al.* (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*.

Kapeli, K. *et al.* (2016). Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. *Nature communications*, **7**, 12143.

Kazan, H. *et al.* (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS computational biology*, **6**, e1000832.

Kent, W. J. *et al.* (2002). The human genome browser at UCSC. *Genome research*, **12**(6), 996–1006.

König, J. *et al.* (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, **17**(7), 909–915.

Licatalosi, D. D. *et al.* (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**(7221), 464–469.

Lin, C. and Miles, W. O. (2019). Beyond CLIP: advances and opportunities to measure RBP-RNA and RNA-RNA interactions. *Nucleic acids research*, **47**(11), 5490–5501.

Maticzka, D. *et al.* (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.*, **15**(1), R17.

Meyer, C. *et al.* (2018). The TIA1 RNA-Binding protein family regulates EIF2AK2-Mediated stress response and cell cycle progression. *Molecular cell*, **69**(4), 622–635.e6.

Pan, X. and Shen, H.-B. (2018). Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, **34**(20), 3427–3436.

Pan, X. *et al.* (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC genomics*, **19**(1), 511.

Pan, X. *et al.* (2019). Recent methodology progress of deep learning for RNA-protein interaction prediction. *Wiley Interdisciplinary Reviews: RNA*, **10**(6), 3627.

Pan, X. *et al.* (2020). RBPsuite: RNA-protein binding sites prediction suite based on deep learning. *BMC genomics*, **21**(1), 884.

Rao, R. *et al.* (2019). Evaluating protein transfer learning with tape. *Advances in Neural Information Processing Systems*, **32**.

Rao, R. *et al.* (2021). Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*.

Rives, A. *et al.* (2020). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*.

Rodriguez, J. M. *et al.* (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic acids research*, **41**(Database issue), D110–7.

Rogers, A. *et al.* (2020). A primer in BERTology: What we know about how BERT works. *arXiv*, page 2002.12327.

Sato, K. *et al.* (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature communications*, **12**(1), 941.

Stražar, M. *et al.* (2016). Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, **32**(10), 1527–1535.

Tahir, M. *et al.* (2021). kDeepBind: Prediction of RNA-Proteins binding sites using convolution neural network and k-gram features. *Chemometrics and Intelligent Laboratory Systems*, **208**, 104217.

Uhl, M. *et al.* (2020). GraphProt2: A novel deep learning-based method for predicting binding sites of RNA-binding proteins. *bioRxiv*, page 850024.

Van Nostrand, E. L. *et al.* (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods*, **13**(6), 508–514.

Van Nostrand, E. L. *et al.* (2020a). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**(7818), 711–719.

Van Nostrand, E. L. *et al.* (2020b). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome biology*, **21**(1), 90.

Vaswani, A. *et al.* (2017). Attention is all you need. *arXiv*, page 1706.03762.

Vig, J. and Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Vig, J. *et al.* (2021). BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*.

Wang, Z. *et al.* (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS biology*, **8**(10), e1000530.

Yan, J. and Zhu, M. (2020). A review about RNA–Protein-Binding sites prediction based on deep learning. *IEEE Access*, **8**, 150929–150944.

Yan, Z. *et al.* (2019). Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*, **35**(14), i333–i342.

Yan, Z. *et al.* (2020). Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. *Bioinformatics*, **36**(Supplement_1), i276–i284.

Yates, A. D. *et al.* (2020). Ensembl 2020. *Nucleic acids research*, **48**(D1), D682–D688.

Yu, H. *et al.* (2019). beRBP: binding estimation for human RNA-binding proteins. *Nucleic acids research*, **47**(5), e26.

Zhang, H. *et al.* (2020). LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics*, **36**(Supplement_1), i258–i267.

Zhang, S.-W. *et al.* (2019). Prediction of the RBP binding sites on lncRNAs using the high-order nucleotide encoding convolutional neural network. *Analytical biochemistry*, **583**, 113364.