Briefings in Bioinformatics, 2022, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

ABLNCPP: Attention Mechanism-based Bidirectional LSTM for non-coding RNA Coding Potential Prediction

First Author,[1,*] Second Author,[2] Third Author,[3] Fourth Author[3] and Fifth Author[4]

[1]School of Computer Science and Engineering, Central South University, 410075, Changsha, China, [2]Department, Organization, Street, Postcode, State, Country, [3]Department, Organization, Street, Postcode, State, Country and [4]Department, Organization, Street, Postcode, State, Country

[*]Corresponding author. email-id.com

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

With the rocketing progress of ribosome profiling, sequencing technology and proteomics, evidence is mounting that non-coding RNA(ncRNA) may be a novel source of peptides or proteins. These peptides and proteins play crucial roles in inhibiting tumor progression, interfering with cancer metabolism and other essential physiological processes. However, existing studies perform well in classifying ncRNAs and mRNAs, but no research has been explicitly raised to distinguish whether ncRNA transcripts have coding potential. For this reason, we propose a bidirectional LSTM network based on attention mechanism called ABLNCPP to assess coding possibility of ncRNA sequences. Moreover, since the embedding method of previously developed classification models CPC2, CPPred and DeepCPP ignores the sequential formation, we introduce a novel non-overlapping trinucleotide(NOLT) coding method for ncRNA to obtain the embeddings of ncRNAs containing sequential characteristics. The extensive evaluations on datasets show that ABLNCPP outperforms other state-of-the-art methods. Taken together, ABLNCPP overcomes the bottleneck of ncRNA coding potential prediction and is expected to provide valuable contributions to cancer discovery and treatment in the future. Source code and datasets can be available at https://github.com/jiangying/ABLNCPP.

Key words: non-coding RNAs, coding potential prediction, sequential characteristics, bidirectional LSTM

Introduction

Next-generation high throughput transcriptome sequencing technology has generated large volumes of novel transcripts in various species. Only 1-2% of these transcribed genes encode proteins; the rest are transcribed as non-coding RNAs[1]. The term non-coding RNA (ncRNA) is generally employed for RNA transcribed from a genome but has no ability to encode a protein. However, with the rapid development of ribosome profiling, sequencing technology and proteomics, increasing discoveries have indicated that ncRNA may have coding potential. Peptides or proteins encoded by these ncRNAs may have some correlations with cancer. Up to now, ncRNAs with known functions include miRNAs, circRNAs, lncRNAs and so on. For example, HOXB-AS3[2] is one of the long non-coding RNAs, and it encodes a conserved 53-aa small peptide that is capable of suppressing colon cancer growth effectively. FBXW7-185aa[3] encoded by circular RNA circ-FBXW7 can inhibit proliferation and induce cell cycle arrest in glioma cells. In CaSki cervical carcinoma cells, specific disruption of circE7[4] is conducive to preventing the translation of the E7 oncoprotein. Thus, exploring whether an ncRNA sequence has the coding probability is of great significance in many aspects such as cancer discovery and treatment.

A variety of researches related to identifying ncRNAs and mRNAs have been developed. While traditional, biological characteristics-based RNA classification methods are restricted by current scientific knowledge, computational methods such as machine learning can detect the complex intrinsic information of transcriptome independently to provide assistance for the sequence coding ability prediction[5]. Therefore, an increasing number of scholars integrate genomics-related knowledge into computational approaches to predict the coding potential of RNA transcripts and have achieved encouraging results. These methods are mainly divided into two categories: (1)Alignment-based approaches; (2)Alignment-free approaches.

Sequence alignment methods heavily rely on the information of transcriptome in the already known databases. This approach is incredibly effective for regulatory RNAs and protein-coding genes which are highly conservative[6]. For example, Kong et al.[7] proposed a classifier based on a support vector machine (SVM), which is called Coding Potential Calculator (CPC). CPC extracts six features with biological meanings from RNA nucleotide sequence, such as log-odds

score, frame score, number of hits and so on, to highly accurately differentiate coding and non-coding transcripts. PhyloCSF[8] took advantage of multiple alignments in a phylogenetic framework and analyzed nucleotide sequence alignment of multiple species. And then, PhyloCSF calculated meaningful likelihood between two phylogenetic models of coding and non-coding genes to identify whether an RNA transcript contains a sub-sequence of the protein-coding region.

However, due to several limitations, such as lineage specific[9], quality of alignments[10], time-consuming and so on, the alignment-based approaches are hard to apply to novel transcripts. In contrast, the alignment-free method mainly benefits from the intrinsic information of the sequence, it is more flexible than the method relying on sequence alignment. Therefore, scholars have fed related data into machine learning to obtain better effects. For example, Coding Potential Assessment Tool(CPAT)[6], mainly constructed a logical regression model with four internal features to identify coding and noncoding transcripts from plenty of candidates, i.e. fickett TESTCODE statistic, open reading frame(ORF) size, ORF coverage and hexamer usage bias. CPC2[11] is an upgraded version of CPC[7] for higher accuracy and efficiency. CPC2 is special-neutral and is more feasible for non-model transcriptomes, which also exhibits superior performance especially on long non-coding transcripts. PLEK[12] combined the SVM algorithm with the improved K-mer scheme to differentiate lncRNAs from mRNAs without utilizing genomic sequences or annotations. Considering the three models mentioned above performed poorly in distinguishing between small mRNAs and small ncRNAs, Tong et al.[13] proposed an approach called CPPred, which means Coding Potential Prediction. This model integrated the commonly used sequence features into SVM, such as ORF integrity, FICKETT score, Hexamer score, isoelectric point and so on. In particular, 30 new features called CTD which contain the composition, transition and distribution information of nucleotide as global transcript sequence descriptors have been added in CPPred. FEElnc[14] utilized random forest to annotate lncRNAs accurately even in the absence of ncRNAs training set.

In addition, deep learning and its variants have been gradually applied to various fields and have shown superior performance. Deep learning is also conducive to discovering intricate hidden structures in high-dimensional data. Zhang et al.[15] proposed DeepCPP, a Convolutional Neural Network utilizing discontinuous k-mer, minimum distribution similarity and nucleotide bias as features. The high performance of mRNA RNN (mRNN)[5] has shown that gated RNNs are able to extract complex information in full-length human transcripts even training with small amount of data and no prior concept of what features define messenger RNAs. DeepRibo proposed by Clauwaert et al.[16] combined features extracted from binding site transcript patterns and ribosome profiling information with convolutional layers and recurrent memory cells. And it is served as a precise tool for identifying ORF in prokaryotes without prior knowledge of the translational landscape. RNAsamba[17] is a deep learning model which utilized IGLOO to process whole transcripts and ORFs. It can identify coding signals in local ORFs and UTR sequences, which is evidence of RNAsamba is independent of the presence of whole coding regions.

Although the studies above exhibit noticeable results, they also have certain limitations. The embedding approach of previously developed classification models CPC2, CPPred and DeepCPP ignores the internal sequential information, which may be useful to the prediction performance to a large extent. To address this issue, we introduce a novel non-overlapping trinucleotide(NOLT) embedding method to obtain the representations of ncRNA transcripts containing sequential characteristics. Moreover, as we mentioned at the beginning, a growing number of ncRNAs have been proven to have coding probability. However, the existing researches identify all ncRNAs as non-coding, which is obviously unreasonable. Since there is no suitable model, we aim to develop an effective coding potential prediction tool named ABLNCPP especially for ncRNAs. ABLNCPP takes advantage of NOLT embedding method and other discriminative features and constructs the architecture based on bidirectional LSTM and attention mechanism. We evaluate ABLNCPP and compare it with five state-of-the-art classification models, ABLNCPP has shown prominent performance and has improved on all indicators. The main advantages of ABLNCPP are as follows:

- As far as we know, ABLNCPP is the first coding potential prediction model specifically designed for ncRNA in this field. It eliminates the limitations of existing models that classified all ncRNAs as having no coding ability. In addition, the ABLNCPP also has prediction function for accessing the coding probability of a novel sequence never seen before to provide a reference for the ncRNA function authentication.
- Being differ from previous researches, we employ a novel embedding approach in ABLNCPP. No matter machine learning or deep learning studies, they all take features calculated by particular formulas as inputs, such as hexamer score, CTD, fickett score and so on, which may ignore critical sequential information within transcripts. To prevent this problem, we introduce a novel NOLT embedding method for capturing key characteristics. Since this method considers all codon composition conditions, it contributes more to classification problem than that of other embedding approaches, such as k-mer nucleotide composition features, one-hot vector and so on.
- Most deep learning models for coding potential prediction utilize Convolutional Neural Network (CNN) to construct main network structure. In ABLNCPP, we propose a novel network that is mainly applied to text classification[18], relation classification[19] and sentiment classification[20]. Although the attention-based bidirectional LSTM has never been used for RNA sequence-related problems, it is conducive to achieving superior model performance.
- Considering that vast amount of studies related to non-coding RNA only took lncRNA as the training subject, ignored that other types of ncRNAs may also have research significance. Even though relevant data on other types of ncRNAs are scarce, our study contains several ncRNAs other than lncRNAs to make the research more comprehensive, such as misc-RNAs, sRNAs, rRNAs and so on.

Materials and Methods

Data Set

The data sets of ABLNCPP consist of two primary databases: SPENCER[21] and LNCipedia 5[22]. SPENCER authenticated 29526 ncRNA-encoded small peptides(ncPEPs) across 15 cancer types through MS-based proteomics analysis pipeline. These ncPEPs are translated from 6803 ncRNA sequences

collected from RNAcentral including 6803 lncRNAs, 34 misc-RNAs, 2 rRNAs, 2 sRNAs and 1 snRNA. And 22060 of the ncPEPs were experimentally validated in other researches. We regard the 6842 ncRNA sequences as positive samples.

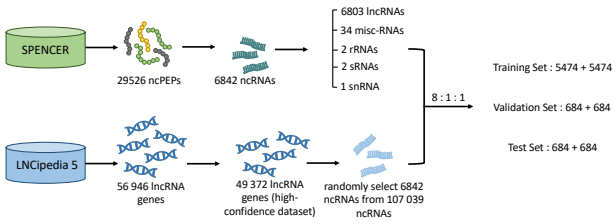

Fig. 1. The flowchart of building training set, validation set and testing set. The positive samples derive from SPENCER and the negative samples randomly selected according to positive samples derive from LNCipedia 5.

LNCipedia was first released in 2012[23], it is a database that contains 21 488 annotated lncRNA sequences of Homo sapiens from different data sources. LNCipedia merged redundant transcripts and grouped transcripts into genes to generate a highly consistent database. In LNCipedia 5, several new lncRNAs were appended into the database benefitting from the release of valuable resources such as FANTOM CAT[24]. Besides, an improved filtering pipeline was introduced to eliminate transcripts derived from protein coding genes. And transcripts that have exons overlapping with coding ORFs will not be regarded as lncRNAs. By this mean, 9203 genes were removed and 455 novel genes were added compared to the previous version, which brought the amount of lncRNA genes in LNCipedia 5 to 56 946 and 127 802 transcripts. A subset of LNCipedia 5 called the high-confidence set includes 49 372 genes and 107 039 transcripts lacking coding potential by any metric. Hence, it is suitable for taking the high-confidence set as negative samples.

Since the amount of high-confidence transcripts in LNCipedia 5 is 107 039, far more than the corresponding number of positive examples. In order to avoid the imbalance of datasets, we select part of transcripts randomly in the high-confidence dataset according to the SPENCER database with fewer data. The ratio of dividing training, test and validation set is 8 : 1: 1. Specifically, we take 5474, 684, 684 transcripts in SPENCER and high-confidence dataset as training set, test set and validation set respectively. The flowchart of building datasets of ABLNCPP can be seen in Figure 1.

Model Framework

tokenization: We used a modified k-mer approach to analyze RNA sequences by breaking them down into non-overlapping segments, removing the first base and repeating the process to obtain a total of k sequences for analysis.

Sequence Embedding: We create sequence embedding by utilizing word embedding techniques.

nBAT (ncRNA BiLSTM-Attention Transformer): By combining BiLSTM and Transformer Encoder, the model has improved performance on ncRNA sequence tasks, better capturing long-term dependencies and learning intrinsic features of ncRNA transcripts.

Fig. 2. By using this approach, we are able to condense long sequences while maintaining the important information.

Feature Learning: Then, we use Transformer Encoder to extract the intrinsic features of the ncRNA transcripts. Transformer Encoder learns the interdependencies between different nucleotide combinations in the ncRNA transcripts and extracts valuable information.

classifier training: We train a multi-layer perceptron (MLP) classifier with the header labels to predict the coding potential of the ncRNA transcripts.

model evaluation: We evaluate the performance of the model on a test dataset. We use common metrics such as accuracy, precision, and recall to measure the model's prediction accuracy.

Tokenization

In our research, we employed a modified k-mer approach to analyze RNA sequences. This method involves breaking down the sequence into non-overlapping segments of length k, where k is the selected k-mer size. To gain further insights, we removed the first base and segmented the sequence again. This process was repeated k times, discarding any segments that were shorter than k bases. As a result, we obtained a total of k sequences for analysis, as illustrated in Figure 2.

Sequence Embedding

In this paper, we employ a PyTorch implementation of word embeddings for sequence embeddings in the task of predicting the coding potential of non-coding RNA (ncRNA). To handle the long sequences, we use average pooling and max pooling on the embeddings to reduce their dimensionality and speed up the training of the downstream model, while having minimal impact on the final prediction results. We also incorporated a learnable position encoding to enhance the model's ability to understand the context and relationships between ncRNA. This position encoding captures the relative position of ncRNA in a sentence, making it particularly useful for sequence prediction tasks. By training the position encoding along with other model parameters, we were able to achieve improved overall performance.

Proposed Architecture

Our model is trained for a binary task and includes intrinsic features, as depicted in Figure 3.

To extract the intrinsic characteristics of ncRNA, we have added an Intrinsic Features Transformer Layer after the xth layer in the encoder. This layer includes an attention head that is specifically trained for this purpose.

BiLSTM-Attention Transformer

The Transformer model[25] is known for its powerful abilities in sequence modeling, including the ability to automatically

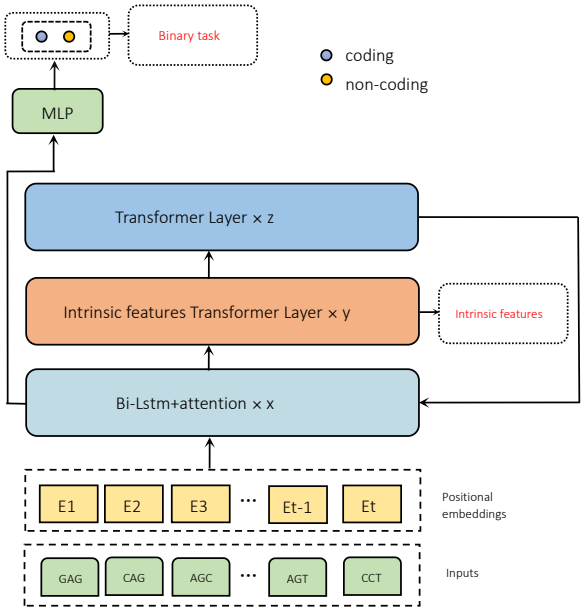

Fig. 3. The proposed architecture is a multi-layered system, in which the number of layers (x, y and z) can vary depending on the specific implementation. The key components of the architecture include a multilayer perceptron (MLP) and the ability to vary the number of layers to optimize performance.

learn long-term dependencies and generate high-quality representations. However, it may still be limited in certain tasks that require long-term dependencies. By combining an LSTM layer with the Transformer model, we can leverage the strengths of both models to capture long-term dependencies and historical information in the input sequence, resulting in improved sequence modeling and generation. This approach is particularly useful for addressing ncRNA-related problems, as the Transformer model alone may struggle with dependencies in ncRNA sequences, while the LSTM model excels in this area. By connecting an LSTM layer before the input of the Transformer model, the model can utilize the memory capabilities of LSTM to better capture long-term dependencies in the input sequence. This can help the Transformer model to better understand the context and meaning of the input, leading to more accurate and coherent output. Additionally, the LSTM layer can also help the model to identify and eliminate noise or irrelevant information in the input, which can improve the efficiency and performance of the Transformer model. One of the key advantages of connecting an LSTM layer after the output of the Transformer model is that it allows the model to re-utilize the memory capabilities of LSTM on the basis of the Transformer's high-quality representation. Additionally, the LSTM layer can correct any errors in dependencies captured by the Transformer and map the high-dimensional output to a lower-dimensional space, making the model's output more compact. This approach can also improve the model's ability to handle ncRNA-related tasks by better capturing dependencies in ncRNA sequences. Furthermore, by using the attention mechanism of LSTM, the model can focus on the most important parts of the output, which can further improve the performance of the model.

The transformer encoder consists of a series of encoder layers, each of which is composed of a multi-head self-attention mechanism and a feedforward network. The self-attention mechanism allows the encoder to attend to different parts of the input sequence at each step, enabling it to capture long-range dependencies and relationships in the data. The feedforward network applies a non-linear transformation to the input, allowing the encoder to capture more complex patterns in the data.

The attention mechanism in transformer encoders can be represented using the following formula:

$$F_h^{(i)} = softmax\left(\frac{Q_h^{(i)} {K_h^{(i)}}^T}{\sqrt{d_k}}\right) V_h^{(i)} \tag{1}$$

In this formula, $F_h^{(i)}$ is the output of the attention mechanism for the i^{th} encoder layer and the h^{th} attention head.

The attention mechanism takes in three input matrices: the query matrix $Q_h^{(i)}$, the key matrix $K_h^{(i)}$, and the value matrix $V_h^{(i)}$. These matrices are usually the outputs of linear transformations applied to the input data. The attention mechanism computes the dot product between the query and key matrices, scaled by the inverse of the square root of the dimensionality of the keys (d_k). The resulting matrix is passed through a softmax function, which normalizes it so that the values in the resulting matrix sum to 1. The resulting matrix is then multiplied by the value matrix $V_h^{(i)}$ to produce the final output of the attention mechanism.

Here is the formula for the final output of the multi-head attention mechanism for the i^{th} encoder layer in a transformer encoder:

$$Multi^{(i)} = [F_1^{(i)}, F_2^{(i)}, \cdots, F_H^{(i)}]W^F \tag{2}$$

In this formula, $Multi^{(i)}$ is the final output of the multi-head attention mechanism, $F_1^{(i)}, F_2^{(i)}, \cdots, F_H^{(i)}$ are the outputs of the attention mechanism for each of the H attention heads, and W^F is a learnable weight matrix.

The attention mechanism in transformer encoders consists of multiple attention heads, each with its own query, key, and value matrices. The outputs of these multiple attention heads are concatenated and passed through a linear transformation represented by the matrix W^F to produce the final output of the multi-head attention mechanism. This allows the transformer encoder to capture more complex patterns and relationships in the input data.

The forward and backward LSTM equations are as follows:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{4}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{5}$$

$$g_t = \tanh(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{7}$$

$$h_t = o_t \odot \tanh(c_t) \tag{8}$$

Where i_t, f_t, o_t, g_t, c_t, and h_t are the input gate, forget gate, output gate, input modulation gate, cell state, and hidden state, respectively. W_{ix}, W_{fx}, W_{ox}, W_{gx}, W_{ih}, W_{fh}, W_{oh}, W_{gh} are the weight matrices for the inputs, x_t, and the previous hidden states, h_{t-1}. b_i, b_f, b_o, b_g are bias vectors. σ is the sigmoid function and $\odot$ denotes element-wise multiplication.

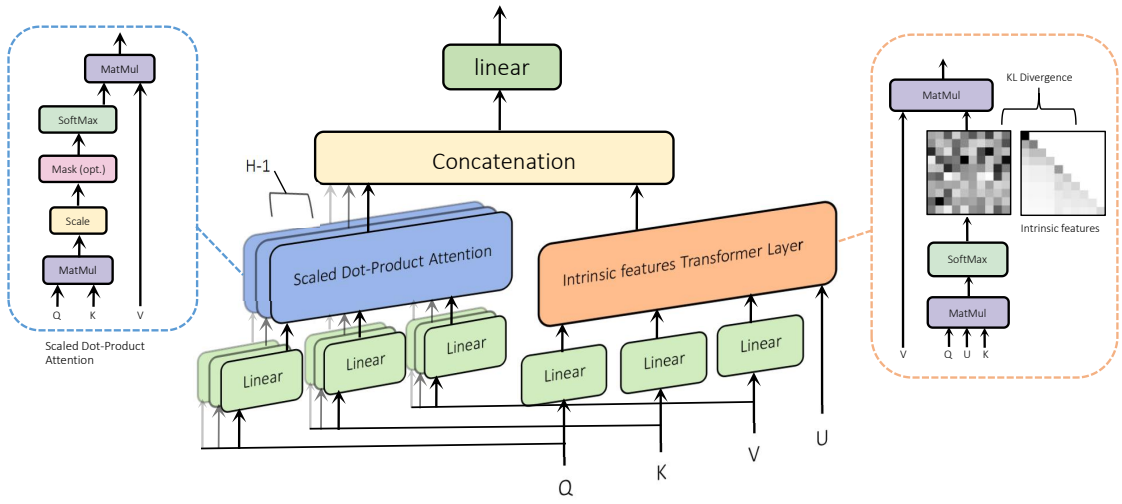

Fig. 4. The Transformer layer has an attention head specifically trained to identify the intrinsic features of non-coding RNA (ncRNA). For each ncRNA, this attention head generates a probability distribution over all positions in the sequence, indicating the strength of the intrinsic characteristic association at each position. The loss function for this model is calculated as the average Kullback-Leibler (KL) divergence between the output distributions of the ncRNAs and their corresponding prior distributions.

For a bidirectional LSTM, the output is the concatenation of the hidden states of the forward and backward LSTMs:

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{9}$$

Similar to the attention mechanism in transformer encoders, the attention mechanism in our BiLSTM layer is represented by the following formula:

$$e_m^{(k)} = H_t^T \, softmax(H_t h_t) \tag{10}$$

In our proposed architecture, we use a multilayer perceptron (MLP) classifier on the output of the last Bilstm+attention layer to classify the embeddings.

The loss function employed in our model is cross-entropy, which is applied to the output of the last BiLSTM+attention layer. The loss on a single utterance is calculated as follows:

$$p_m^{(k)} = MLP(e_m^{(k)}) \tag{11}$$

$$\mathcal{L}^{seq} = -\sum_{s=1}^{S} y_s^{bin} \log(p_m^{(k)}) \tag{12}$$

Where $p_m^{(k)}$ represents the predicted probability of the ith ncRNA sequence being a particular label, S is the total number of label categories, and y_s^{bin} is the one-hot representation of the ground truth ncRNA sequence label.

Intrinsic Features

The intrinsic features transformer encoder layer, shown in Figure 4, is a modified version of the standard Transformer encoder layer. It includes an attention head that is trained to predict the degree of intrinsic characteristic association between each non-coding RNA (ncRNA) in the input sequence. Later, we will demonstrate the effectiveness of this approach in the Results section.

The standard attention head in the Transformer encoder layer generates an output matrix of size T x T, where T is the length of the input sequence. Each row of this matrix contains attention weights indicating the importance of each token in the input sequence for a given token. In contrast, the intrinsic features attention head is trained to assign weights only to the features governors (i.e., ancestors) of each ncRNA, resulting in an output matrix of the same size.

To train the intrinsic features attention head, we define a loss function based on the difference between the output attention weight matrix of the intrinsic features attention head and a predefined prior attention weight matrix. The prior matrix encodes the prior knowledge that each non-coding RNA (ncRNA) should attend to its feature extraction ancestors, with higher attention weights assigned to ancestors that are closer to the ncRNA. During training, we obtain the prior attention weights using the output of the Feature Extraction part of the model. The goal of training is to minimize the loss by adjusting the attention weights of the intrinsic features attention head to match those of the prior matrix.

To train the attention head to focus on the intrinsic features of each ncRNA, we define prior attention weights for each ncRNA based on the distance between the ncRNA and its intrinsic features. Specifically, the prior attention weights for ncRNA j are defined as follows:

$$w_{i,j}^{prior} = \begin{cases} softmax(-d_{i,j}/\tau) & \text{if } i <= j \\ 0 & \text{if } i > j \end{cases} \tag{13}$$

In the equations above, $d_{i,j}$ is the distance between token i and j, $softmax$ is the Softmax function, and τ is the temperature parameter controlling the variance of the attention weights over all the intrinsic features and $i, j \in \{1, 2, \cdots, T\}$ the values of i and j are intrinsic features of the corresponding ncRNA. The stack of prior attention weights for all T ncRNAs is a $T \times T$ matrix denoted by W^{prior}.

The goal of training is to minimize the difference between W^{prior} and the attention matrix $W_h^{(s)}$ output by the attention head h at the sth layer. This difference is measured using the mean row-wise Kullback-Leibler (KL) divergence between these two matrices, which is used as an additional loss function in

addition to the binary task loss functions. We refer to this loss as the dependency loss, denoted by $\mathcal{L}^{lin}$. Formally:

$$\mathcal{L}^{fea} = \frac{1}{T}\sum_{j=1}^{T} D_{KL}(W_j^{prior}||W_{h,j}^{(s)}) \tag{14}$$

$$W_h^{(s)} = softmax(Q^{(s)}U^{lin\cdot}(K^{(s)})^T) \tag{15}$$

In these equations, $D_{KL}(\cdot)$ denotes the KL divergence, $Q^{(s)}$ and $K^{(s)}$ are linear transformations of $E(s-1)$, $W_{h,j}^{(s)}$ is the jth row of $W_h^{(s)}$, and $U^{lin\cdot}$ is a parameter matrix. In equation (5), we use the biaffine attention instead of the scaled dot product attention, which has been shown to be effective[26].

We treat τ as a hyperparameter and tune it on the validation set. With $\tau \rightarrow 0^+$ the attention head will be trained to only pay attention to the direct parent of each token, our method is a general approach compared to [27].

Multi-task Learning

We utilize a multi-task learning approach[28] to train our model. Our loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}^{seq} + c^{fea} \cdot \mathcal{L}^{fea} \tag{16}$$

Where c^{fea} are considered as hyperparameters and are chosen based on their performance on the validation set.

Feature Extraction

In order to mine the information from different perspectives to represent the raw ncRNA sequence into vectors in a more comprehensive way, we select 10 features of sequence content and homology as auxiliaries. The sequence embedding vectors extracted by BiLSTM based on attention mechanism are connected with features to feed into the linear layer for prediction.

The features we utilized can be divided into two categories: one is intrinsic features of ncRNA transcripts, which can be derived from sequence directly and are widely used in classification problems related to coding potential; and the other represents the properties of ncRNA transcripts.

The intrinsic features consist of 5 components: ORF length, ORF coverage, ORF integrity, Fickett Score and Hexamer Score. As increasing researches have proved that ncRNAs contain ORFs, we introduce three features related to ORF, i.e. ORF length, ORF coverage, and ORF integrity. ORF length is the max length of the ORF detected in three forward frames. ORF coverage is defined as the ratio of ORF length to ncRNA sequence length and is complementary to ORF length to a great extent. ORF integrity indicates whether an ORF region is complete, in other words, the ORF region begins with a start codon(ATG) and ends with a stop codon(TAG, TAA, or TGA). Since a long, integrated and putative ORF is difficult to be observed in ncRNA without coding ability, ORF related features are the most discriminative features and the most fundamental criteria to identify whether ncRNAs can encode proteins.

Fickett Score reflects the nucleotide composition and codon usage bias of ncRNA transcript by calculating nucleotide composition and position frequency. The nucleotide composition is the percentage of each base contributing to the sequence. The position frequency is an indicator of the degree to which base is favored in one codon position compared with another.

$$\begin{aligned} A_1 &= \text{ The number of As in position } 0, 3, 6 \ldots \\ A_2 &= \text{ The number of As in position } 1, 4, 7 \ldots \\ A_3 &= \text{ The number of As in position } 2, 5, 8 \ldots \end{aligned}$$

$$A_{pos} = \frac{\max(A_1, A_2, A_3)}{\min(A_1, A_2, A_3) + 1} \tag{17}$$

The calculation of position values of other nucleotides(T, C, G) is the same as above. The Fickett Score is computed as follows:

$$\text{Fickett Score } = \sum_{i=1}^{8} p_i w_i \tag{18}$$

where the p_i is obtained by a look up table which converts the eight values into TESTCODE scores, and w_i is the weight for respective base which indicate the percentage of time for every parameter predicts the coding type of a transcript alone successfully.

Hexamer Score is one of the most discriminative characteristics to measure differential hexamer usage bias between ncRNA with and without coding potential. For a given hexamer sequence $\mathrm{H} = S_1, S_2, \ldots, S_n$:

$$\text{Hexamer Score } = \frac{1}{n}\sum_{i=1}^{n} \log(\frac{F(S_i)}{F'(S_i)}) \tag{19}$$

where $F(S_i)$ and $F'(S_i)$ represent in-frame frequency($i = 1, 2, \ldots, 4096$) calculated from positive and negative datasets respectively.

Properties of ncRNA transcript we used are first proposed by LncFinder[29], including Peak, signal-to-noise ratio(SNR) and three kinds of distances. Each nucleotide of non-coding RNA sequence, i.e. A, T, C, G, has an EIIP value, indicating the energy of delocalized elections in nucleotides[30]. EIIP can effectively avoid the potential bias caused by the speculated translation process compared with the value of pI. The $X[i]$ represents the EIIP indicator of ncRNA sequence, the $X[k]$ and the power spectrum$P[k](k = 0, 1, 2, \ldots, L-1)$ can be calculated as follows:

$$X[k] = \sum_{i=0}^{L-1} X[i]e^{-j\frac{2\pi ki}{L}}, \mathrm{P}[k] = |\mathrm{X}[k]|^2 \tag{20}$$

The feature Peak is the signal at 1/3 position and can be represented as the formula:

$$Peak = P[\frac{L}{3}] = |X[\frac{L}{3}]|^2 \tag{21}$$

The $\bar{E}$ is the average power, so that the SNR can be obtained by the following formula:

$$\bar{E} = \frac{\sum_{k=0}^{L-1} P[k]}{L}, \mathrm{SNR} = \frac{\mathrm{X}\left[\frac{\mathrm{L}}{3}\right]}{\bar{\mathrm{E}}} \tag{22}$$

Finally, we employ the Logarithm-distance to ncRNA(Dist.NC), Logarithm-distance to protein-coding transcript(Dist.PCT) and distance ratio(Dist.Ratio) to further quantify the usage bias of hexamer.

Implementation Details

In this study, we implemented our experiments using PyTorch [31]. The hyperparameters were chosen based on the performance on the validation set. We utilized the Adam optimizer[32] with $\beta_1 = 0$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$. A weight decay factor, as described in[33], was also applied. The learning rate schedule began by increasing the learning rate linearly from 0 to 0.00005, a process known as warming up, and then decreased according to the values of a cosine function. The number of warming up steps, approximately 20% of the total training steps, was determined by validation performance. We utilized the Transformers library[34] for the implementation of the optimizer and learning rate scheduler. A dropout rate of 0.2[35] were used. The models were trained for 100 epochs using a batch size of 128. The best testing results were obtained from the checkpoints with the highest validation performance. The hidden dimension of the Transformer encoder layer was set to 768, and the size of the feedforward layer was 3072. The input dimension for the BiLSTM layer is 200 and the hidden layer dimension is 768. The value of c^{fea} is set to 0.5 in our model. Our model features two Transformer encoder layers, each equipped with 4 attention heads, and one BiLSTM layer, which has 1 attention head. (as shown in Figure 2 with x = 1,y = 1 and z = 0). The intrinsic features are computed by considering the position of tokens represented by i and j. The variables i and k represent the attention of transformer and bi-directional LSTM respectively, while h and m represent the layer of transformer and bi-directional LSTM respectively.

Performance Evaluation

We compute the wildly utilized standard performance metrics which are the accuracy(ACC), precision(PRE), sensitivity(SN), specificity(SP), F-score, Matthews correlation coefficient(MCC) and AUC to measure ABLNCPP performance. These evaluation indexs are defined as follows:

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Precision(PRE) = \frac{TP}{TP + FP}$$

$$Sensitivity(SN) = \frac{TP}{TP + FN}$$

$$Specificity(SP) = \frac{TN}{TN + FP}$$

$$F - score = \frac{2 * PRE * SN}{PRE + SN}$$

$$Matthews\ correlation\ coefficient(MCC) = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TP + FP) * (TN + FP) * (TN + FN)}}$$

where TP stands for true positive, FN, TN, FP represent false negative, true negative and false positive, respectively. The MCC is an overall measurement of binary classification problem and an objective assessment index. AUC is the value of the area under the receiver operating characteristic curve.

Results and Discussion

We conduct five parts of experiments, i.e., biological feature selection, embedding visualization, ablation experiments, contrastive experiments and case study. In feature selection, mRMR technology is used to screen the features with the strongest correlation with the classification performance to generate the final characteristics subset. Moreover, we adopt t-SNE to visualize the effect of ABLNCPP for classifying a massive volume of transcripts in the second part. The ablation of network structure is designed in ablation experiments. In addition, we compare the performance of ABLNCPP with the most representative coding potential prediction models, including CPAT, PLEK and CPC2 which show excellent performance in ncRNA and mRNA classification, as well as models using deep learning, such as CPPred and RNAsamba. Finally, case study is carried out to verify the effectiveness of ABLNCPP.

Feature Selection

Table 1. Feature Order of mRMR

Order	Feature
1	ORF length
2	ORF integrity
3	OFR coverage
4	Peak
5	Fickett Score
6	SNR
7	Hexamer Score
8	Dist.NC
9	Dist.PCT
10	Dist.Ratio

Since there may exist some noise information in these representative features which may be detrimental to model performance, we use mRMR tool to select appropriate characteristics. First, the mutual information among features and between each feature and category to be predicted will be calculated utilizing the following formula:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{23}$$

And then, we make use of the average value of the mutual information between each feature and category to obtain the correlation $C(F,c)$:

$$C(F,c) = \frac{1}{|F|} \sum_{f_i \in F} I(f_i;c) \tag{24}$$

where F represents feature set and c represents class of prediction. The average value of the mutual information among features will be utilized to calculate the redundancy $R(F)$:

$$R(F) = \frac{1}{|F|^2} \sum_{f_i, f_j \in F} I(f_i;f_j) \tag{25}$$

where f_i and f_j represent features in the feature set. The final result mRMR is the trade-off of correlation and redundancy:

$$mRMR = \max_F \left[\frac{1}{|F|} \sum_{f_i \in F} I(f_i;c) - \frac{1}{|F|^2} \sum_{f_i, f_j \in F} I(f_i;f_j) \right] \tag{26}$$

We rank the 10 features in Table 1. In order to further verify the performance variation of the model after adding

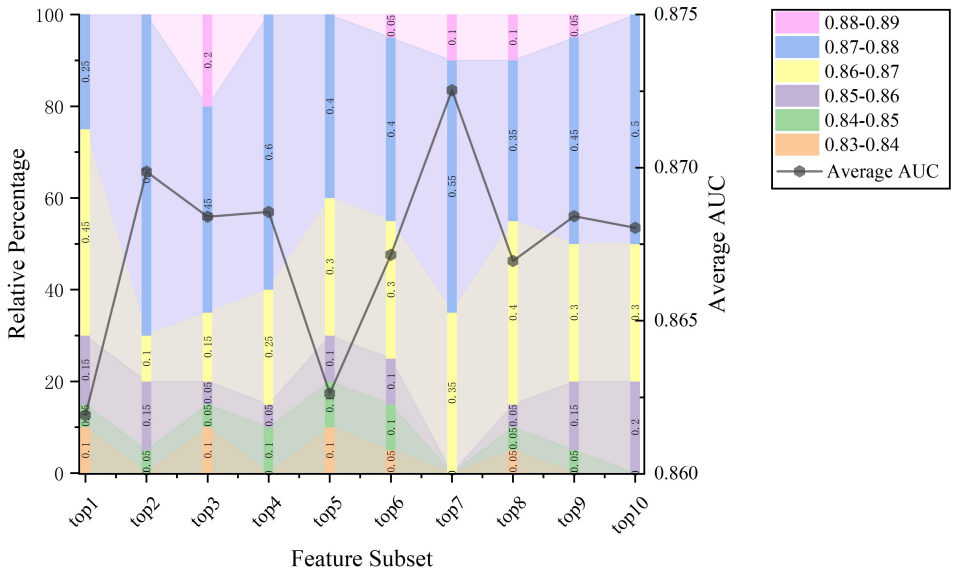

Fig. 5. The ABLNCPP performance of different feature subsets. Each color in the graph represents a section of AUC value, the numbers on the bar chart represent relative percentages, and each point on the line chart represents the average AUC value.

features, we conduct experiments on each feature to be added into ABLNCPP by the order obtained by mRMR. In particular, we carry out 20 times experiments on each combination and calculate the average value of AUC. And the result is illustrated in Figure 5. It can be seen that the AUC value of the first 7 features are almost distributed in the range of 0.86 to 0.87 and 0.87 to 0.88. With the assistance of the first 7 features, the model performance is the most prominent which achieves 0.872 and others are all below 0.87. Thus, we determine to regard the first 7 features as the final characteristics subset.

Embedding Visualization

We further compare the difference in embedding distribution before and after the classification through t-SNE, an unsupervised dimensionality technique that can be used to visualize embedding of ncRNA sequences. In Figure 6, each dot represents a sample, red represents samples with no coding potential and gray denotes samples with coding potential. As seen from the figure, samples before classification are almost overlapping. After the BiLSTM based on attention mechanism in ABLNCPP, the two types of samples have an apparent tendency to converge in different directions. Although few part of the embeddings in the middle is relatively close, most sample points are gathered at both ends. Generally speaking, ABLNCPP can effectively distinguish the differences between the subset with and without coding potential.

Ablation Experiments

After determining the features, we carry out ablation experiments on the network structure. The results are exhibited in details in Figure 7. To detect the effectiveness of our encoding approach, we compare the subsequence embedding(represented in Figure 7 as SE) method in DeepLncLoc with our method. As seen in the figure, the AUC of the previously proposed method is nearly 2.1% lower than NOLT embedding method. The MCC of our approach is much more higher than it is in DeepLncLoc, approximately 6.7%. Other evaluation indicators of subsequence embedding method are generally lower than ours. It is fully proved that our method can extract the potential internal information of sequence more comprehensively by exploring the information in head and tail junctions of subsequences.

In addition, we delete the auxiliary features for comparison and only feed the obtained embedding of ncRNA transcript into the model. The result shows that classification performance is severely influenced, which exactly indicates that adding features essentially contributes to our model for supplementing the information about ORF, the properties of transcripts and so on. Moreover, we modify the network structure to use only unidirectional LSTM instead of bidirectional LSTM, or remove the attention mechanism to investigate the validity of ABLNCPP. Although the SN of structure without attention mechanism and the SP of unidirectional LSTM is almost identical with ABLNCPP, the whole performance of them is worse. By contrast, the BiLSTM combined with attention mechanism can significantly take effect in our prediction task.

Contrastive Experiments

As the first proposed coding potential prediction tool designed explicitly for ncRNAs, we present the prediction accuracies of our ABLNCPP, as well as the performance comparison with current representative models in ncRNA and mRNA classification tasks, such as CPC2, CPAT and PLEK. CPPred is a newly proposed classification model and outperforms previous methods on sORF type coding and noncoding RNAs. RNAsamba makes use of a deep learning model IGLOO that processes the whole transcript and ORF to identify ncRNAs and mRNAs. To benchmark the models, it is fair to retrain all programs on the same training set and test against the same test set. Since PLEK does not deal with sequences

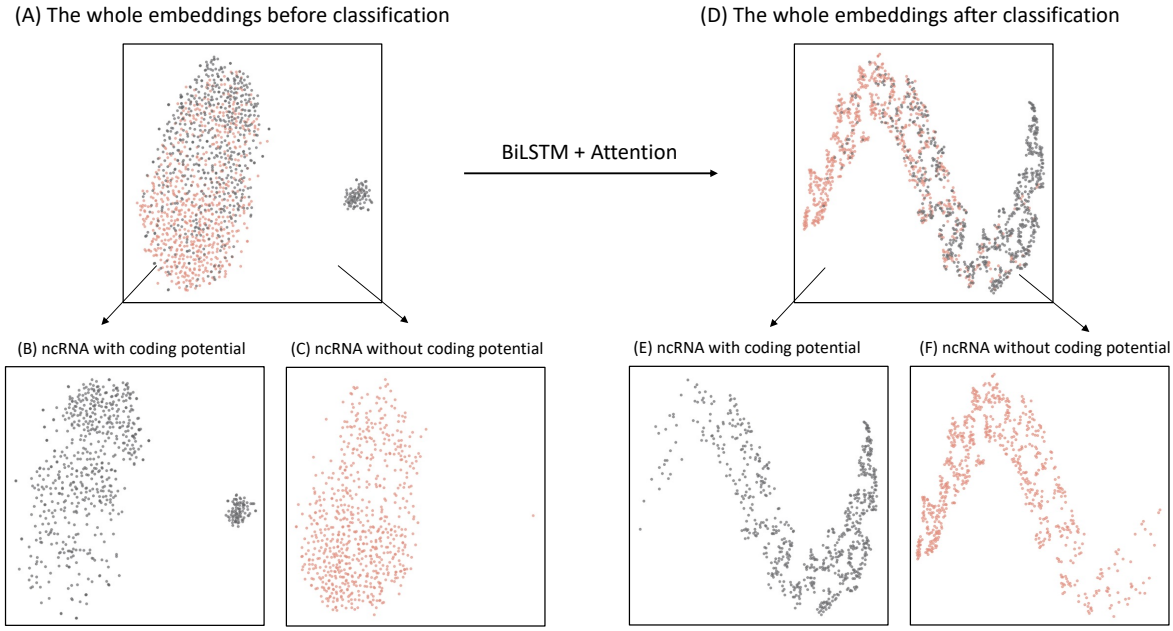

Fig. 6. Visualization of sequence embedding. The (A)-(C) are the embedding distribution before classification, and the (D)-(F) are the embedding distribution after classification. (A) and (D) are the whole samples, (B) and (E) are the positive samples, (C) and (F) are the negative samples.

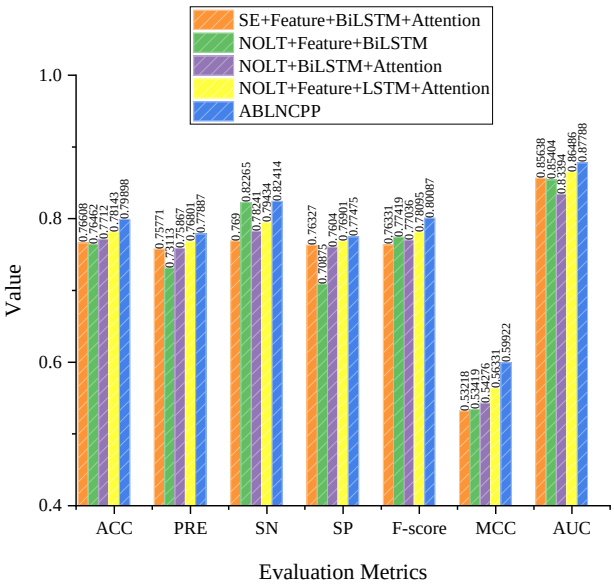

Fig. 7. The performance of structure ablation experiments.

whose lengths are smaller than 201, we conduct an objective comparison on the filter version of our datasets by only keeping ncRNAs with lengths longer than 200, and the results are illustrated in Table 2.

ABLNCPP achieves sensitivity of 82.41% and F-score of 0.801, meanwhile the AUC of 87.8%. Compared with other tools on the same test sets, ABLNCPP has the highest value of all kinds of measurements. Taking the best indicator of all models as a comparison, the ACC is improved nearly 2.8% from 77.12% to 79.90%, the PRE is improved almost 1.1% from 76.73% to 77.89%, the SN achieves above 80% from 79.24% to 82.41%, the SP is enhanced nearly 1.1% from 76.39% to 77.47%, the F-score achieve above 0.80 from 0.773 to 0.801, the MCC is improved almost 5.7% from 0.542 to 0.599 and AUC increase nearly 2.3% from 0.855 to 0.878. In total, based on the observations with limited data amount above, ABLNCPP shows superiority to CPAT, CPC2, PLEK, CPPred and RNAsamba, which certainly has more application significance for processing actual data.

Case Study

We conduct a case study on several experimentally verified non-coding RNA transcripts which not exist in the training, validation, and testing sets of our model to further confirm the availability of ABLNCPP.

As is shown in table 3, the ncRNAs verified by ABLNCPP are proven to play a vital role in inhibiting the spread of cancer. AKT3[36] inhibits glioblastoma tumorigenicity by competing with active phosphoinositide-dependent Kinase-1. An 87-amino-acid peptide encoded by LINC-PINT[2] directly interacts with polymerase associated factor complex (PAF1c) and suppresses glioblastoma cell proliferation in vitro and in vivo. SHPRH-146aa[37] is a novel protein produced by circRNA SHPRH and the overexpression of it in U251 and U373 glioblastoma cells reduces their malignant behavior and tumorigenicity. PPP1R12A[38] generates circPPP1R12A-73aa which promotes the growth and metastasis of Colon Cancer(CC) via activating Hippo-YAP signaling pathway. The peptide translated by Lgr4[39] interacts with and activates Lgr4, which further activates the Wnt/β-catenin signaling pathway, promoting the self-renewal and tumorigenesis of CC stem cells. Therefore, drugs targeting the Lgr4-encoded peptide may be used to treat CC. A 218aa protein generated by FNDC3B[40] inhibits the expression of Snail, and subsequently promotes the tumor-suppressive effect of FBP1 in CC.

Table 2. ABLNCPP performance comparing with state-of-the-art models

	ACC(%)	PRE(%)	SN(%)	SP(%)	F-score	MCC	AUC
CPC2	70.08	72.20	65.40	74.76	0.686	0.403	0.759
CPAT	73.43	70.99	79.24	67.62	0.749	0.472	0.812
PLEK	74.19	73.90	74.76	73.61	0.743	0.484	0.817
CPPred	75.85	74.98	75.60	76.09	0.753	0.517	0.841
RNAsamba	77.12	76.73	77.85	76.39	0.773	0.542	0.855
ABLNCPP	**79.90**	**77.89**	**82.41**	**77.47**	**0.801**	**0.599**	**0.878**

Table 3. Case Study

Name	species	Prediction Result	Pubmed ID
AKT3	Homo sapiens	Coding	PMID: 31470874
LINC-PINT	Homo sapiens	Coding	PMID: 30367041
SHPRH	Homo sapiens	Coding	PMID: 29343848
PPP1R12A	Homo sapiens	Coding	PMID: 30925892
Lgr4	Homo sapiens	Coding	PMID: 31269234
FNDC3B	Homo sapiens	Coding	PMID: 32241279
SclA	Drosophila melanogaster	Coding	PMID: 23970561
SclB	Drosophila melanogaster	Coding	PMID: 23970561
CACNA1G-AS1	Homo sapiens	Noncoding	PMID: 30908634
PIK3CD-AS2	Homo sapiens	Noncoding	PMID: 32165621
PRR7-AS1	Homo sapiens	Noncoding	PMID: 32165621

Although we trained our model on human data sets, it also showed generality and stability in predicting the coding potential of transcripts on other species. For example, SclA and SclB[41] are peptides produced by lncRNA which regulate calcium transport hence influence regular muscle contraction, in the Drosophila heart.

Even though lots of ncRNAs lack of coding potential by any metric, they also influence vital physiological processes such as evolution, heritable variation and disease treatment in human. CACNA1G-AS1[42] facilitates hepatocellular carcinoma progression through the miR-2392/C1orf61 pathway. Lung adenocarcinoma progression can be promoted by PIK3CD-AS2[43] via YBX1-mediated suppression of p53 pathway. PRR7-AS1[44] is a biomarker that could be utilized to predict the prognosis of HCC patients and is linked to the infiltration of immune cells in HCC. Thus, ABLNCPP can predict which ncRNAs are possible to encode peptides or proteins and the peptide and proteins produced by these ncRNAs may have an inhibitory effect on cancer. Moreover,it is worth noting that the ncRNAs without coding potential maybe also regulate the cancer development by hindering the signaling pathway.

Conclusion

In this work, we developed a novel feature representation approach called non-overlapping trinucleotide embedding method to learn internal sequential characteristics of ncRNA transcripts comprehensively. It contained all possible conditions of codon composition, and was proved to be effective in identifying ncRNAs with at least 2.1% higher value in AUC than existing methods, and other metrics were all better than before. Additionally, we found that the three features related to ORF and Hexamer Score were discriminative for our tasks, they emerged notable effects in the feature selection experiment. Finally, based on features selected by mRMR and embeddings obtained by NOLT method, together with BiLSTM and attention mechanism, we proposed ABLNCPP for ncRNA coding potential prediction.

T-SNE was applied to visualize the classification effect of ABLNCPP, and it indicated that our model had superior performance in distinguishing between the ncRNA transcripts with and without coding potential. As for network structure, the BiLSTM and attention mechanism had different degrees of contribution, and both of them were indispensable. ABLNCPP and contrastive models were evaluated on the same datasets which filtered out sequences less than 201, the results showed that ABLNCPP outperformed all models, with at least 2.3% improvement on AUC than that of other methods. In a nutshell, ABLNCPP is the first proposed model for ncRNA coding potential prediction and represented prominent performance on all metrics. It is expected to provide valuable contributions to cancer discovery and treatment in the future.

Competing interests

No competing interest is declared.

Author contributions statement

Must include all authors, identified by initials, for example: S.R. and D.A. conceived the experiment(s), S.R. conducted the experiment(s), S.R. and D.A. analysed the results. S.R. and D.A. wrote and reviewed the manuscript.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work is supported in part by funds from the National Science Foundation (NSF: # 1636933 and # 1920920).

References

1. Manel Esteller. Non-coding rnas in human disease. *Nature reviews genetics*, 12(12):861–874, 2011.
2. Jin-Zhou Huang, Min Chen, DE Chen, Xing-Cheng Gao, Song Zhu, Hongyang Huang, Min Hu, Huifang Zhu, and Guang-Rong Yan. A peptide encoded by a putative lncrna hoxb-as3 suppresses colon cancer growth. *Molecular cell*, 68(1):171–184, 2017.
3. Yibing Yang, Xinya Gao, Maolei Zhang, Sheng Yan, Chengjun Sun, Feizhe Xiao, Nunu Huang, Xuesong Yang, Kun Zhao, Huangkai Zhou, et al. Novel role of fbxw7 circular rna in repressing glioma tumorigenesis. *JNCI: Journal of the National Cancer Institute*, 110(3):304–315, 2018.
4. Jiawei Zhao, Eunice E Lee, Jiwoong Kim, Rong Yang, Bahir Chamseddin, Chunyang Ni, Elona Gusho, Yang Xie, Cheng-Ming Chiang, Michael Buszczak, et al. Transforming activity of an oncoprotein-encoding circular rna from human papillomavirus. *Nature communications*, 10(1):1–12, 2019.
5. Steven T Hill, Rachael Kuintzle, Amy Teegarden, Erich Merrill III, Padideh Danaee, and David A Hendrix. A deep recurrent neural network discovers complex biological rules to decipher rna protein-coding potential. *Nucleic acids research*, 46(16):8105–8113, 2018.
6. Liguo Wang, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic acids research*, 41(6):e74–e74, 2013.
7. Lei Kong, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research*, 35(suppl_2):W345–W349, 2007.
8. Michael F Lin, Irwin Jungreis, and Manolis Kellis. Phylocsf: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, 2011.
9. Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927, 2011.
10. Patrick D Schloss. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16s rrna gene-based studies. *PLoS computational biology*, 6(7):e1000844, 2010.
11. Yu-Jian Kang, De-Chang Yang, Lei Kong, Mei Hou, Yu-Qi Meng, Liping Wei, and Ge Gao. Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, 45(W1):W12–W16, 2017.
12. Aimin Li, Junying Zhang, and Zhongyin Zhou. Plek: a tool for predicting long non-coding rnas and messenger rnas based on an improved k-mer scheme. *BMC bioinformatics*, 15(1):1–10, 2014.
13. Xiaoxue Tong and Shiyong Liu. Cppred: coding potential prediction based on the global description of rna sequence. *Nucleic acids research*, 47(8):e43–e43, 2019.
14. Valentin Wucher, Fabrice Legeai, Benoit Hedan, Guillaume Rizk, Lætitia Lagoutte, Tosso Leeb, Vidhya Jagannathan, Edouard Cadieu, Audrey David, Hannes Lohi, et al. Feelnc: a tool for long non-coding rna annotation and its application to the dog transcriptome. *Nucleic acids research*, 45(8):e57–e57, 2017.
15. Yu Zhang, Cangzhi Jia, Melissa Jane Fullwood, and Chee Keong Kwoh. Deepcpp: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for rna coding potential prediction. *Briefings in bioinformatics*, 22(2):2073–2084, 2021.
16. Jim Clauwaert, Gerben Menschaert, and Willem Waegeman. Deepribo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic acids research*, 47(6):e36–e36, 2019.
17. Antonio P Camargo, Vsevolod Sourkov, Gonçalo A G Pereira, and Marcelo F Carazzolle. Rnasamba: neural network-based assessment of the protein-coding potential of rna sequences. *NAR genomics and bioinformatics*, 2(1):lqz024, 2020.
18. Gang Liu and Jiabao Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338, 2019.
19. Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212, 2016.
20. Weijiang Li, Fang Qi, Ming Tang, and Zhengtao Yu. Bidirectional lstm with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing*, 387:63–77, 2020.
21. Xiaotong Luo, Yuantai Huang, Huiqin Li, Yihai Luo, Zhixiang Zuo, Jian Ren, and Yubin Xie. Spencer: a comprehensive database for small peptides encoded by noncoding rnas in cancer patients. *Nucleic acids research*, 50(D1):D1373–D1381, 2022.
22. Pieter-Jan Volders, Jasper Anckaert, Kenneth Verheggen, Justine Nuytens, Lennart Martens, Pieter Mestdagh, and Jo Vandesompele. Lncipedia 5: towards a reference set of human long non-coding rnas. *Nucleic acids research*, 47(D1):D135–D139, 2019.
23. Pieter-Jan Volders, Kenny Helsens, Xiaowei Wang, Björn Menten, Lennart Martens, Kris Gevaert, Jo Vandesompele, and Pieter Mestdagh. Lncipedia: a database for annotated human lncrna transcript sequences and structures. *Nucleic acids research*, 41(D1):D246–D251, 2013.
24. Chung-Chau Hon, Jordan A Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen JL Rackham, Julian Gough, Elena Denisenko, Sebastian Schmeier, Thomas M Poulsen, Jessica Severin, et al. An atlas of human long non-coding rnas with accurate 5 ends. *Nature*, 543(7644):199–204, 2017.
25. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
26. Timothy Dozat and Christopher D Manning. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*, 2016.
27. Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*, 2018.

28. R Caruana. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Citeseer, 1993.
29. Siyu Han, Yanchun Liang, Qin Ma, Yangyi Xu, Yu Zhang, Wei Du, Cankun Wang, and Ying Li. Lncfinder: an integrated platform for long non-coding rna identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Briefings in bioinformatics*, 20(6):2009–2027, 2019.
30. Achuthsankar S Nair and Sivarama Pillai Sreenadhan. A coding measure scheme employing electron-ion interaction pseudopotential (eiip). *Bioinformation*, 1(6):197, 2006.
31. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
32. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
33. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
34. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
35. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
36. Xin Xia, Xixi Li, Fanying Li, Xujia Wu, Maolei Zhang, Huangkai Zhou, Nunu Huang, Xuesong Yang, Feizhe Xiao, Dawei Liu, et al. A novel tumor suppressor protein encoded by circular akt3 rna inhibits glioblastoma tumorigenicity by competing with active phosphoinositide-dependent kinase-1. *Molecular cancer*, 18(1):1–16, 2019.
37. Maolei Zhang, Nunu Huang, Xuesong Yang, Jingyan Luo, Sheng Yan, Feizhe Xiao, Wenping Chen, Xinya Gao, Kun Zhao, Huangkai Zhou, et al. A novel protein encoded by the circular form of the shprh gene suppresses glioma tumorigenesis. *Oncogene*, 37(13):1805–1814, 2018.
38. Xiao Zheng, Lujun Chen, You Zhou, Qi Wang, Zhuojun Zheng, Bin Xu, Chen Wu, Qi Zhou, Wenwei Hu, Changping Wu, et al. A novel protein encoded by a circular rna circppp1r12a promotes tumor pathogenesis and metastasis of colon cancer via hippo-yap signaling. *Molecular cancer*, 18(1):1–13, 2019.
39. Xiaofei Zhi, Jiaxuan Zhang, Zhouyang Cheng, Longjun Bian, and Jun Qin. Circlgr4 drives colorectal tumorigenesis and invasion through lgr4-targeting peptide. *International journal of cancer*, 2019.
40. Zihao Pan, Jianye Cai, Jiatong Lin, Huinian Zhou, Jingwen Peng, Jinliang Liang, Long Xia, Qi Yin, Baojia Zou, Jun Zheng, et al. A novel protein encoded by circfndc3b inhibits tumor progression and emt through regulating snail in colon cancer. *Molecular cancer*, 19(1):1–15, 2020.
41. Emile G Magny, Jose Ignacio Pueyo, Frances MG Pearl, Miguel Angel Cespedes, Jeremy E Niven, Sarah A Bishop, and Juan Pablo Couso. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, 341(6150):1116–1120, 2013.
42. Jinghui Yang, Chunsheng Li, Hang Li, and Changyong E. Lncrna cacna1g-as1 facilitates hepatocellular carcinoma progression through the mir-2392/c1orf61 pathway. *Journal of Cellular Physiology*, 234(10):18415–18422, 2019.
43. Xiufen Zheng, Junying Zhang, Tian Fang, Xiaoxiao Wang, Siwei Wang, Zhifei Ma, Youtao Xu, Chencheng Han, Mengting Sun, Lin Xu, et al. The long non-coding rna pik3cd-as2 promotes lung adenocarcinoma progression via ybx1-mediated suppression of p53 pathway. *Oncogenesis*, 9(3):1–17, 2020.
44. Yifan Lu, Songhai Chen, Qingqing Wang, Jie Zhang, and Xuanzeng Pei. Prr7-as1 correlates with immune cell infiltration and is a diagnostic and prognostic marker for hepatocellular carcinoma. *Journal of Oncology*, 2022, 2022.

Author Name. This is sample author biography text. The values provided in the optional argument are meant for sample purposes. There is no need to include the width and height of an image in the optional argument for live articles. This is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text.

Author Name. This is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text.