이학석사 학위논문

Long Noncoding RNA Identification using Recurrent Neural Network

순환신경망을 이용한 lncRNA 판별

2017 년 8 월

서울대학교 대학원

협동과정 생물정보학 전공

백 정 환

이학석사 학위논문

Long Noncoding RNA Identification using Recurrent Neural Network

순환신경망을 이용한 lncRNA 판별

2017 년 8 월

서울대학교 대학원

협동과정 생물정보학 전공

백 정 환

Long Noncoding RNA Identification using Recurrent Neural Network

지도교수 윤 성 로

이 논문을 이학석사 학위논문으로 제출함

2016 년 11 월

서울대학교 대학원

협동과정 생물정보학 전공

백 정 환

백정환의 이학석사 학위논문을 인준함

2016 년 12 월

위 원 장	김 선	(인)
부위원장	윤 성 로	(인)
위 원	천 종 식	(인)

Abstract

Long Noncoding RNA Identification using Recurrent Neural Network

Junghwan Baek

Interdiciplinary Program in Bioinformatics

College of Natural Sciences

The Graduate School

Seoul National University

Long noncoding RNAs (lncRNAs) are important regulatory elements in biological processes. LncRNAs share similar sequence characteristics with messenger RNAs (mRNAs), but they play completely different roles, thus providing novel insights for biological studies. The development of next-generation sequencing (NGS) has helped in the discovery of lncRNA transcripts. However, the experimental verification of numerous transcriptomes is time consuming and costly. To alleviate these issues, a computational approach is needed to distinguish lncRNAs from the transcriptomes.

We present a deep learning-based approach, lncRNAnet, to identify lncRNAs that incorporates recurrent neural networks (RNNs) for RNA sequence modeling and convolutional neural networks (CNNs) for detecting stop codons to obtain an open reading frame (ORF) indicator. LncRNAnet performed clearly better than the other tools for sequences of short lengths, on which most lncRNAs are distributed. In addition, lncRNAnet successfully learned features and showed 7.83%, 5.76%, 5.30%, and 3.78%

improvements over the alternatives on a human test set (HT) in terms of specificity, accuracy, F1-score, and area under the curve (AUC), respectively.

Keywords: Long noncoding RNA (lncRNA), Recurrent Neural Network (RNN), Deep Learning

Student Number: 2015-20507

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Only less than 2% of the three billion base pairs in the human genome encode proteins [1–4], and the functions of the remaining parts remain unknown. Among these remaining parts, noncoding RNAs (ncRNAs), which refer to transcripts that are not translated into proteins, are often considered key regions responsible for various biological processes.

The long noncoding RNAs (lncRNAs), which are ncRNAs composed of more than 200 nucleotides (nt), are of particular interest. Although the low conservation of lncRNAs often makes them appear as transcriptional noise [5, 6], the known lncRNAs play key roles as regulatory elements in biological processes and engage in various disease processes, including cancers, neurological disorders, and immunological disorders [7, 8]. Identifying lncRNAs is thus essential for understanding gene regulation and the potential causes of important diseases.

Typically, lncRNAs have sequence characteristics similar to those of protein-coding transcripts, making lncRNAs difficult to identify, whereas short ncRNAs (sncRNAs) are clearly distinguished from protein-coding transcripts. Some lncRNAs even undergo

transcriptional and post-transcriptional processes just like protein-coding transcripts [9]. Although various transcriptomes have been sequenced to date, many lncRNAs remain undiscovered and exist as "dark regions" in the genome [10]. In expression measurements, the low expression levels of lncRNAs are treated as anomalies, hindering their discovery [11, 12]. The application of next-generation sequencing (NGS) has improved the efforts to annotate lncRNAs in terms of cost and accuracy, but many experimental approaches to verifying lncRNA transcriptomes still require significant time and resources.

To complement available experimental techniques, computational lncRNA identification methods have been proposed [13–23]. Their common formulation is to use a binary classifier that can predict whether a given nucleotide sequence encodes an lncRNA. Various classification models have been used, including support vector machines (SVMs), random forests (RFs), and neural networks.

Existing computational methods heavily rely on the features extracted from identified lncRNAs and/or their comparative genomics-based profiles obtained by database searches and multiple sequence alignments (MSAs). The use of the features and MSA profiles may be necessary to some extent to reveal some common patterns of lncRNAs. However, it clearly limits the accuracy and robustness of lncRNA identification and the opportunity to identify novel ncRNAs that have subtle properties different from those of the known lncRNAs.

Specifically, existing techniques often suffer from the following issues. First, DB searches are useful for highly conserved sequences (e.g., protein-coding transcripts) but not for lncRNAs. They have lower conservation values than the exons of the protein-coding transcripts. Consequently, DB search-based identifications of lncRNAs may erroneously predict lncRNAs as coding transcripts [24]. In addition, DB searches are limited to well-annotated species and otherwise produce unsatisfactory results, and the underlying alignment operations are often time consuming and affected by

alignment parameters that are usually set heuristically. Second, many of the previous computational approaches depend on manually crafted features and heuristic decision criteria. Certain features may helpful in designing a robust lncRNA classifier, but identifying and validating effective features require a substantial amount of human effort. Even with such effort, manually designed features may still fail to capture non-canonical signals that exist in elusive lncRNAs. The use of an unprincipled approach for determining decision criteria can hurt the accuracy and generalization (i.e., the performance of a machine learning algorithm for the data not used for training) of the classifier.

To overcome these limitations, we propose lncRNAnet, a deep learning-based approach for identifying lncRNAs. The key characteristics and notable contributions of our approach include the following:

1. The proposed lncRNAnet successfully learned intrinsic features by incorporating recurrent neural networks (RNNs) for RNA sequence modeling.

2. An open reading frame (ORF) indicator was proposed by exploiting stop codon detections based on convolutional neural networks (CNNs). The ORF indicator adopts prior knowledge about translation to reinforce the model.

3. LncRNAnet successfully detected short lncRNA candidates and robustly performed in the global length range, which is important because lncRNAs are relatively shorter than mRNAs due to smaller exon numbers.

To validate our approach, we tested lncRNAnet with two datasets containing 7,000 transcripts each and compared it with four existing tools in terms of widely used metrics. In our experiments, the proposed lncRNAnet successfully learned the inherent features of lncRNA transcripts and delivered the highest performance, outperforming the best alternative on the HT dataset in terms of specificity, accuracy, F1-score, and the area

under the curve (AUC) by 7.83%, 5.76%, 5.30%, and 3.78%, respectively.

Chapter 2

Background

As mentioned in the previous section, the existing approaches use various handcrafted features to identify lncRNAs. The novelty of our method is the use of deep CNNs and RNNs to learn the features of lncRNAs. A sequence of nucleotides can be divided into a set of consecutive non-overlapping triplets (*i.e.*, six reading frames). Among the reading frames, an open reading frame (ORF) is a series of codons that have the potential to be translated. The lncRNAs can be considered a complementary set of coding transcripts that includes the ORFs; hence, it is essential to detect the ORFs to distinguish lncRNAs from the abundant transcripts. Based on this point, we use CNNs to detect the ORFs as the coding transcript candidates and stacked RNNs to distinguish lncRNAs from them.

2.1 Long noncoding RNA (lncRNA)

LncRNAs are an RNA family that does not encode proteins, and they play special roles in various biological processes by regulating gene expression [25]. Specifically, lncRNAs act as *cis* and *trans* elements, regulating nearby and distant genes, respectively,

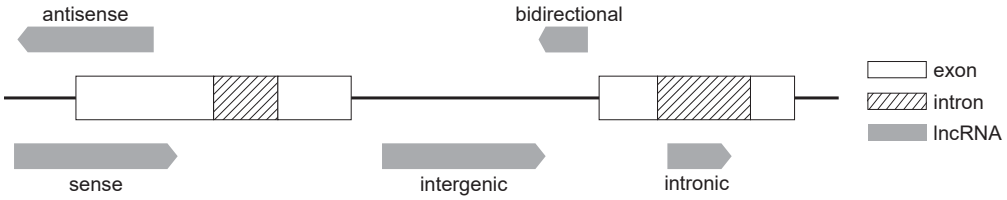

Figure 2.1 Five locations of lncRNAs in genomes
Antisense, sense, bidirectional, intergenic and intronic region. Each block shows the position in the genomic region, exon (white), intron (slash) and lncRNA (grey). The direction of lncRNA shows a direction of the strand.

to activate or degrade the transcription process by binding to the transcriptional factors. The characteristics of lncRNAs are similar to those of protein-coding transcripts, such as splicing, 5′-capping, and poly-A tailing [26–28], although their roles are different. Furthermore, the fact that some lncRNAs act as both protein-coding and noncoding transcripts complicates lncRNA classification [29].

Unlike protein coding mRNAs, lncRNAs coincide in various regions in genomes. From Figure 2.1 it can be categorized into 5 regions. (a) sense, whose sequence is positioned in the sense strand of protein coding gene, which can be aligned to intron/ exon region (b) antisense is the sequence overlapped with an antisense strand of protein coding gene (c) intronic, whose sequence is lied in the intron of transcript (d) intergenic, whose sequence is included between two genes without any overlapping genes and (e) bidirectional, whose sequence is present in opponent sequence distant from the protein coding region [9]. lncRNA in intergenic regions are known as long intergenic noncoding RNA (lincRNA). It is a largest subgroup of lncRNA that were discovered [30]. lincRNAs which does not overlap with protein coding genes are remarked to simplify the analysis of lncRNAs and actively researched [31].

To distinguish lncRNAs from protein-coding transcripts, we can consider the following rules [29]: ORF-based, sequence- and structure-based, and filtering-based rules. The ORF-based rules include the length of an ORF, the conservation of the ORFs, and the

ratio between the length of an ORF and a transcript sequence. The sequence-structure-based rules consider the conservation of the secondary structure. The filtering-based rules preprocess the artifacts caused by the sequencing. The central point of these rules is based on the fact that lncRNAs can be considered as a complementary set of coding transcripts because they do not code proteins.

In this study, we exploit the ORFs detected by CNNs in lieu of the aforementioned handcrafted features.

2.2 Convolutional Neural Network (CNN)

A CNN is a neural network specialized for image data that models the relations of the adjacent pixels. A CNN applies filters in the form of a convolution operation to extract features from the data. The advantage of a CNN is that it reduces the parameters compared to other neural networks by sharing them as multiple filters [32]. The convolution filters share the parameters independent of position; thus, we can reduce the number of used parameters. This parameter sharing of the convolutional filters and the local connections of the nodes increase the performance in handling sparsely connected data. CNNs have shown outstanding performance in two-dimensional sparse data such as images or matrices [33, 34]. In addition, one-dimensional CNNs have been recently applied to sequential data classification, language modeling, and other related problems in natural language processing [35–37].

In this study, we use one-dimensional CNNs to detect the ORFs as the candidates of coding transcripts.

2.3 Recurrent Neural Network (RNN)

An RNN is a neural network that feeds the output of a previous cell as the current input of the network (*i.e.*, acyclic graph). This shape can help the RNNs to learn the

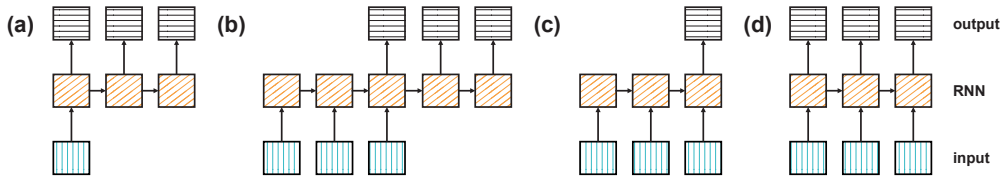

Figure 2.2 Four types of RNN
(a) one-to-many (b) many-to-many with different input-output length (c) many-to-one (d) many-to-many with same input-output length. input is described as vertical lined blue box, RNN is described as diagonal lined box, and output is described as horizontal lined grey box.

sequential behavior; hence, they are widely used for classifying sequences such as in sequence-to-sequence learning and speech recognition. An acyclic graph can be described as a series of connections of cells. This graph can be unrolled infinitely according to the length of the input, and the parameters of each cell are shared through each time step; thus, it behaves flexibly for the sequential data. The training procedure using the unrolled network is called backpropagation through time (BPTT). The BPTT method has constraint regarding training time with inputs of highly variable lengths, so we used a bucketing technique to address this issue.

Although RNNs show outstanding performances on various tasks, they are vulnerable to long sequences in saving long-term dependencies. For a standard RNN, learning the whole input causes the vanishing gradient problem during the training. To control the data flow of the internal memories to learn the long-term dependencies by adding gates, Hochreiter and Schmidhuber [38] developed the long short-term memory (LSTM), and Cho et al. [39] developed the gated recurrent unit (GRU).

The majority of neural networks is fixed input-fixed output type [40]. In contrary, a RNN can behave in several forms corresponding to many input-output combinations. Figure 2.2 (a) is used to fixed input-variable output, which can extract information from stationary input (*i.e.* description of images). Figure 2.2 (c) is a variable input-fixed output combination, which can classify sequential data. Figure 2.2). Figure 2.2 (b) and

(d) is for sequential input- sequential output. Figure 2.2 (b) is applied to numerous natural language processing tasks. Figure 2.2 (d) can perform in simultaneous tasks such as video image recognitions.

In this study, RNNs with LSTM units are used to learn the intrinsic behavior of lncRNAs.

Chapter 3

Proposed methodology

Figure 3.1 shows the overview of our proposed method, lncRNAnet. To determine whether a given sequence is an lncRNA, our method uses the following four phases: (a) bucketing, (b) detecting ORF indicators, (c) encoding sequences, and (d) learning lncRNAs. Algorithm 1 shows the details of each phase. Our model accepts candidate transcript sequences as inputs. In the bucketing phase, each sequence is sorted into buckets with regard to their sequence length (lines 6–8). Then, an ORF indicator is identified for each sequence (line 9). In the encoding sequences phase, the transcript sequence and the ORF indicator are pre-padded to match the maximum sequence length of each bucket. Using a one-hot encoding scheme, the transcript sequence and the ORF indicator are preprocessed (lines 10–13). To learn lncRNAs, the sequence data are trained through the whole architecture (lines 18–29).

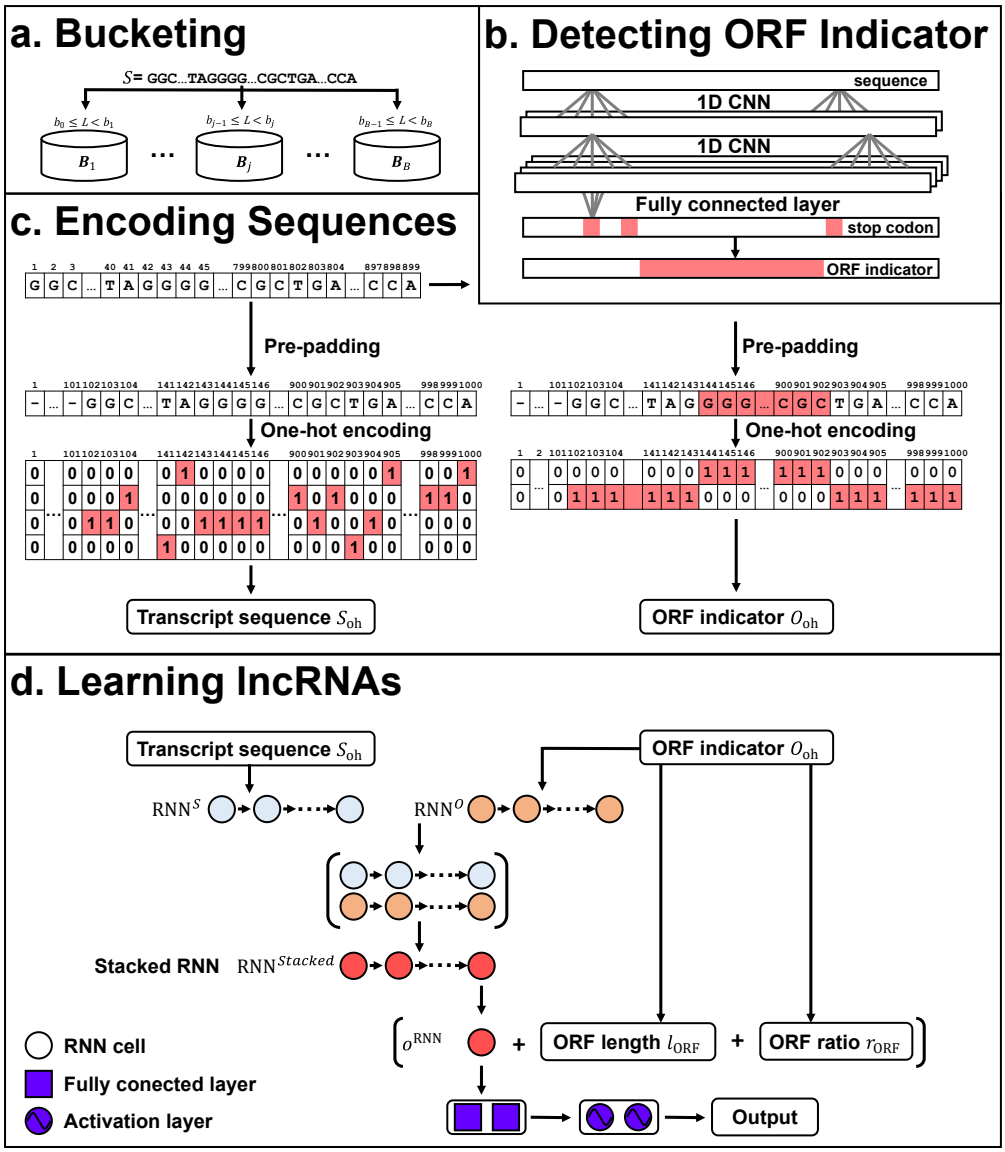

Figure 3.1 The overview of lncRNAnet

First, each sequence is sorted into a bucket, and an ORF indicator is obtained by CNNs. Then, a transcript sequence and an ORF indicator are encoded. The RNN model uses a transcript sequence and an ORF indicator as inputs, and then, an ORF length and an ORF ratio are calculated from the ORF indicator as additional features of the model.

3.1 Bucketing

As mentioned in the previous section, RNNs can expand their shape with regard to the input sequence length. Because the length distribution of transcripts varies from hundreds to hundreds of thousands, we decided to use an RNN as the building block of our architecture. However, if the lengths of the sequences in a dataset vary greatly, randomly selected sequences in a batch will have different lengths. In this case, the length of the sequences can be made to be equal through sequence padding, but sparse values (*e.g.*, [0, 0, 0, 0]) of padded inputs may occupy a large proportion and may hamper training time. Bucketing can alleviate these sparse inputs (*i.e.*, padded values) by selecting batches from a bucket, which contains sequences with similar lengths [41]. A sequence S of length L such that $b_{j-1} < L \leq b_j$ where $j \in \{1, \cdots, B\}$ is sorted into a bucket $\mathbf{B}_j$ with the maximum sequence length b_j. Batch-wise training with sequences of similar lengths can be performed later by padding each sequence in the bucket to the maximum sequence length of the bucket.

3.2 Detecting an ORF Indicator

All protein-coding transcripts possess ORFs, sequences that can be translated. Identifying candidate ORFs in the transcript $S = (s_1, \cdots, s_L)$ is an important guideline for distinguishing lncRNAs from protein-coding transcripts. Normally, ORFs can be identified by finding sequences between a start codon and a stop codon. However, the occurrence of non-canonical start codon signals disturbs the ORF detection. In this study, ORFs are searched using only the stop codons to identify non-canonical signals. ORFs between stop codons are called stop-to-stop ORFs.

Stop-to-stop ORFs are identified using the one-dimensional CNN. The model is shown in Figure 3.1(b). One-dimensional CNNs are stacked, and fully connected layers are attached to detect stop codons for each time step. After identifying all stop codons,

an ORF indicator $O = (o_1, \cdots, o_L)$ is processed by finding the longest stop-to-stop ORF in three forward frames. Each character o_t is encoded to 0 if the character is in the ORF and to 1 if the nucleotide is not in the ORF.

$$o_t = \begin{cases} 0 & s_t \in \text{ORF} \\ 1 & s_t \notin \text{ORF} \end{cases} \quad where \quad t \in \{1, \cdots, L\}$$

From the ORF indicator, we consider two additional ORF-related features, l_{ORF} and r_{ORF}. Let the maximum ORF length be l_{ORF}, and the ratio, l_{ORF} over transcript length L be denoted by r_{ORF}.

3.3 Encoding Sequences

After bucketing, each bucket has sequences with similar lengths. Our method exploits the premise that the RNN unrolls to the maximum sequence length of the batch from $\mathbf{B}_j$, b_j. Hence, sequence padding is mandatory for training. The transcript sequence S is pre-padded with character '$-$' to match the maximum length. The character s_t of sequence S consists of four nucleotides $\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$. A sequence $S = (s_1, \cdots, s_L)$ whose length $b_{j-1} \leq L < b_j$ is pre-padded to S_{pad} as follows:

$$S = (s_1, \cdots, s_L) \xrightarrow{\text{pre-padding}} S_{\text{pad}} = (\overbrace{-, \cdots, -}^{b_j - L}, s_1, \cdots, s_L).$$

The pre-padded sequence S_{pad} is preprocessed to be trained in RNNs. The padded sequence is transformed to S_{oh} by using one-hot encoding [42]. One-hot encoding is a method in which each character in a sequence is represented as a binary vector whose dimension is equal to the number of characters. In this case, nucleotides consist of four characters $\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$. Each nucleotide s_t can be expressed as $\texttt{A} = [1, 0, 0, 0]$, $\texttt{C} = [0, 1, 0, 0]$, $\texttt{G} = [0, 0, 1, 0]$ and $\texttt{T} = [0, 0, 0, 1]$, and the padded value '$-$' can be

represented as $[0, 0, 0, 0]$. For instance, the start codon '`ATG`' can be expressed as a series of four-dimensional binary vectors $\langle [1, 0, 0, 0], [0, 0, 0, 1], [0, 0, 1, 0] \rangle$. As a result, the transcript sequence is projected to a series of four-dimensional tensors of shape (4, b_j).

The ORF indicator sequence O, processed from the transcript sequence S, is also pre-padded with character '$-$' to match the maximum length. The ORF indicator character o_t of O consists of two characters $\{0, 1\}$. The sequence $O = (o_1, \cdots, o_L)$ whose length $b_{j-1} \leq L < b_j$ is pre-padded as follows:

$$O = (o_1, \cdots, o_L) \xrightarrow{\text{pre-padding}} O_{\text{pad}} = (\overbrace{-, \cdots, -}^{b_j - L}, o_1, \cdots, o_L)$$

In addition, O_{pad} is preprocessed to be trained in the RNN by one-hot encoding. As a result, the ORF indicator is projected to a series of two-dimensional tensors of shape (2, b_j).

3.4 Learning lncRNAs

Preprocessed inputs are trained through the neural network. The transcript sequence S_{oh} (dimension (4, b_j)) and the ORF indicator O_{oh} (dimension (2, b_j)) are passed through each masking layer, which ignores the loss that originates from padded values. Thus, the network does not calculate the loss from paddings. Each masked input from the transcript sequence and the ORF indicator is then fed into many-to-many RNNs with n_h cells. As shown in Figure 3.1(d), the output representations of RNN^S and RNN^O for each time step are out_t^S and out_t^O (both having dimension (n_h)). For each time step, outputs out_t^S and out_t^O from two RNNs are merged to be a state x_t^{RNN} of RNN$^{\text{Stacked}}$

(dimension $(2n_h)$),

$$
\begin{aligned}
x^{\text{RNN}} &= \left[x_1^{\text{RNN}}, \cdots, x_t^{\text{RNN}}, \cdots, x_{b_j}^{\text{RNN}}\right] \\
&= \left[\left[\ \text{out}_1^S, \text{out}_1^O\ \right], \cdots, \left[\ \text{out}_t^S, \text{out}_t^O\ \right], \cdots, \left[\ \text{out}_{b_j}^S, \text{out}_{b_j}^O\ \right]\right]
\end{aligned}
$$

where $t \in \{1, \cdots, b_j\}$.

The merged state is sent to a many-to-one stacked RNN, $\text{RNN}^{\text{Stacked}}$, with dropout layers [43]. Dropout layers help reduce the overfitting of the network by dropping connections between perceptrons in the network. Dropping some connections can be seen as a sampling of the proposed network, which generalizes the model. The final output from RNN $\text{out}^{\text{Stacked}}$ is merged with the ORF length l_{ORF} and the ORF ratio r_{ORF} to x^{FC} (dimension $(n_h + 2)$),

$$
x^{\text{FC}} = [\text{out}^{\text{Stacked}}, l_{\text{ORF}}, r_{\text{ORF}}].
$$

The final output is connected to the fully connected layer of dimension two. On the top of the fully connected layer output out^{FC}, each output is out_m^{FC} where $m \in \{0, 1\}$. A softmax activation layer [44] was added to predict binary classification values. A softmax function is expressed as follows, where $\mathbf{y}$ is a label if the candidate is an lncRNA or not.

$$
P(\mathbf{y} = m|\text{out}^{\text{FC}}) = \frac{1/(1 + \exp(-\text{out}_m^{\text{FC}}))}{\sum_{m=0}^{1} 1/(1 + \exp(-\text{out}_m^{\text{FC}}))} \tag{3.1}
$$

In the model training period, from buckets $(\mathbf{B}_1, \cdots, \mathbf{B}_B)$, $\mathbf{B}_k$ is chosen randomly, and the number of sequences equal to the batch size is selected to create batches. Depending on the bucket $\mathbf{B}_k$, the RNNs unroll themselves to their maximum length b_k. For an epoch, it repeats choosing buckets and selecting sequences until all sequences in the dataset are trained to reduce the model loss $\mathcal{L}$. The cross-entropy loss that is

common in binary classification is as follows:

$$\mathcal{L}(\mathbf{y}) = -\frac{1}{n_{\text{batch}}} \sum_{n=1}^{n_{\text{batch}}} (y_n \log(p_n) + (1 - y_n) \log(1 - p_n)) \tag{3.2}$$

where y_n is the label if it is an lncRNA or not, p_n is the probability of a candidate lncRNA transcript, and n_{batch} is the mini-batch size.

Algorithm 1 Pseudocode of lncRNAnet

Input: Sequences $S : s_1, s_2, \cdots, s_L$
▷ $s_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, L: length of S, N: Number of sequences
Output: **y** (lncRNA/ Protein coding transcript)
Maximum sequence length of a bucket : $b_1, b_2, \cdots, b_B$ ▷ $b_0 = 0$
Buckets : $\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_B$ ▷ B: Number of buckets
for $i = 1$ to N **do**
  for $j = 1$ to B **do**
    if $L \leq b_j$ **then**
      $\mathbf{B}_j$.append(S)
      O=ORFindicator(S) ▷ $O : o_1, \cdots, o_L$
      S_{pad}=pre-padding(S)
      O_{pad}=pre-padding(O)
      S_{oh}=one-hot(S_{pad}) ▷ S_{oh}: preprocessed transcript sequence
      O_{oh}=one-hot(O_{pad}) ▷ O_{oh}: preprocessed ORF indicator
      break
    end if
  end for
end for
repeat
  select random number $k \leq B$ to select $\mathbf{B}_k$
  take n_{batch} sequences from $\mathbf{B}_k$ ▷ n_{batch} : mini-batch size
  Calculate ORF length l_{ORF}
  Calculate ORF ratio r_{ORF}
  Compute RNN^S, and RNN^O
  Merge two output representations for each time point out$_t^S$, and out$_t^O$ into x_t^{RNN} ▷ out^S, out^O: output representations of the RNN^S and RNN^S
  Send x_t^{RNN} to stacked RNNs
  Merge out$^{\text{Stacked}}$, l_{ORF} and r_{ORF} as an input to the fully connected layer
  Softmax calculation from the fully connected layer output out$^{\text{FC}}$
  Minimize the cross entropy loss
  $\mathcal{L}(\mathbf{y}) = -\frac{1}{n_{\text{batch}}} \sum_{n=1}^{n_{\text{batch}}} (y_n \log(p_n) + (1 - y_n) \log(1 - p_n))$
until $\#epoch = n_{\text{epoch}}$

Chapter 4

Results

4.1 Datasets

We used human and mouse lncRNA and protein-coding transcript data downloaded from GENCODE [45] release 25. The lncRNAs were used as positive data, and the protein-coding transcripts were used as negative data. Redundant sequences and sequences containing characters other than `A`, `C`, `G`, and `T` were excluded. As shown in Figure 4.1, the lncRNAs less than 3,000 in length cover about 95.22% of the human data and 90.29% of the mouse data. Only sequences shorter than 3,000 were used in our experiments.

The total numbers of human protein-coding transcripts and lncRNAs were 94,127 and 27,692, respectively, and those for the mouse data were 60,474 and 14,226, respec-

Table 4.1 The number of sequences used in our experiments

	Model-Training	HT	MT
lncRNAs	21,000	3,500	3,500
protein-coding transcripts	21,000	3,500	3,500
total	42,000	7,000	7,000

Table 4.2 Effects of hyperparameter variations through a 5-fold cross validation in terms of prediction accuracy

parameter	# of stacked RNN layers (n_l)			# of RNN hidden units (n_h)				dropout probability (p_d)			
	0	**1**	2	25	50	**100**	200	0	0.25	**0.5**	0.75
Training	0.8642	**0.9857**	0.9057	0.8658	0.9545	**0.9857**	0.8760	0.9854	0.9920	**0.9857**	0.9386
Test	0.8410	**0.9018**	0.8575	0.8629	0.8862	**0.9018**	0.8513	0.8924	0.8901	**0.9018**	0.8729

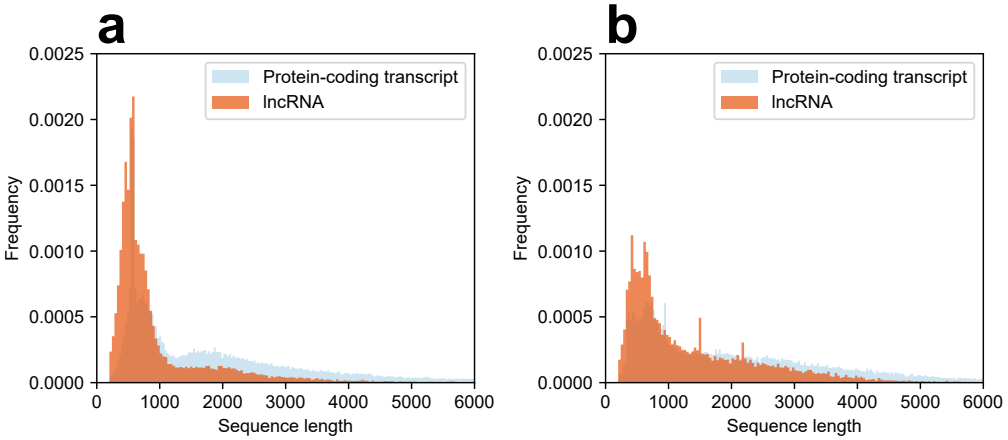

Figure 4.1 Length distribution of the protein-coding transcripts and lncRNAs in the (a) human and (b) mouse data

tively. In our experiment, both the human protein-coding transcripts and lncRNAs were down-sampled to 24,500, where 21,000 were used for training (Model-Training) and 3,500 were used for testing (HT). Those of the mouse sequences were randomly down-sampled to 3,500 only for testing (MT). The details of the numbers of sequences used in our experiments are listed in Table 4.1. Model-Training was used for model selection through a 5-fold cross validation, the HT dataset was used for model generalization for the human data, and the MT was used for cross-species experiments independent of the training species.

4.2 Performance Comparison of Hyperparameter Variations

The empirical optimal hyperparameters were obtained from various combinations based on baseline parameters. The baseline parameters for the number of stacked RNN layers (n_l), the number of RNN hidden units (n_h), and RNN dropout probability (p_d) were 1, 100, and 0.5, respectively. Experiments were performed on Model-Training with a 5-fold cross validation in terms of the average train and test accuracy, and trained by minimizing the categorical cross-entropy with the Adam optimizer [46] for 200 epochs.

Table 4.3 Effects of feature information in terms of prediction accuracy

Features	S, O, l_{ORF},r_{ORF}	S, O	S, l_{ORF}, r_{ORF}	S
Training	**0.9857**	0.9077	0.9668	0.8105
Test	**0.9018**	0.8899	0.8331	0.7204

S: transcript sequence, O: ORF indicator, l_{ORF}: ORF length, r_{ORF}: ORF ratio

Maximum bucket lengths of 500, 1000, 1500, 2000, 2500, and 3000 were used.

Table 4.2 shows the experimental results of the hyperparameter combinations. The number of stacked RNNs layers (n_l) was changed to 0, 1, and 2. The n_l of 0 indicates that the outputs of RNN^S and RNN^O were fed directly into the fully connected layer without additional stacked RNN layers. The n_l of 1, which had one stacked RNN layer, showed the best performance in both training and test accuracy.

The number of RNN hidden units (n_h) was changed to 25, 50, 100, and 200. As the number of RNN hidden units increased, the LSTM memory capacity increased. The n_h of 100 showed the best performance, indicating that the sequence memory capacity of 100 was the most suitable size for distinguishing the lncRNAs from protein-coding transcripts in our experiments.

We changed the dropout probability (p_d) to 0, 0.25, 0.5, and 0.75. The p_d of 0.5 showed the second-highest accuracy in training, but showed the best performance in the generalization test.

Table 4.3 shows the prediction performance according to the certain features. The model with all features, S, O, l_{ORF}, and r_{ORF}, showed the best performance in terms of average training and test accuracy. The model trained with sequences S and O and the model with ORF scalar features, l_{ORF}, and r_{ORF}, followed. The vanilla RNN, which accepted only preprocessed transcript sequences, S, was not trained properly. According to the results, the ORF features, O, l_{ORF}, and r_{ORF}, are crucial feature information for

Table 4.4 Comparison of prediction performance (data: HT)

	Sensitivity	Specificity	Accuracy	F1-score	AUC
lncRNAnet	0.9591	**0.8766**	**0.9179**	**0.9211**	**0.9672**
CPAT	0.9229	0.8129	0.8679	0.8747	0.9264
CNCI	0.9771	0.7214	0.8493	0.8664	0.8205
CPC	**0.9911**	0.5134	0.7523	0.8000	0.9320
PLEK	0.9840	0.5294	0.7567	0.8018	0.9004

Table 4.5 Comparison of prediction performance (data: MT)

	Sensitivity	Specificity	Accuracy	F1-score	AUC
lncRNAnet	0.9463	**0.8903**	**0.9183**	**0.9205**	**0.9667**
CPAT	0.9646	0.8157	0.8901	0.8978	0.9530
CNCI	0.9689	0.7883	0.8786	0.8886	0.8610
CPC	**0.9897**	0.5940	0.7919	0.8262	0.9508
PLEK	0.9180	0.5643	0.7411	0.7800	0.8300

identifying lncRNAs.

4.3 Performance Comparison between Tools

LncRNAnet was compared with the following existing tools: CPC [14], CPAT [15], CNCI [16], and PLEK [18]. The model of lncRNAnet was trained with the 42,000 human dataset Model-Training. The generalization test was performed on the human dataset HT and mouse dataset MT.

4.3.1 Performance Comparison in the Human Dataset

Table 4.4 and Figure 4.3 (a) show the prediction performance on dataset HT. The proposed lncRNAnet outperformed the best alternative results in terms of specificity, accuracy, F1-score, and AUC by 7.83%, 5.76%, 5.30%, and 3.78%, respectively. CPC showed the highest sensitivity, 3.34% higher than that of lncRNAnet, and the second-highest AUC, but it showed the lowest performance in terms of specificity, accuracy, and F1-score.

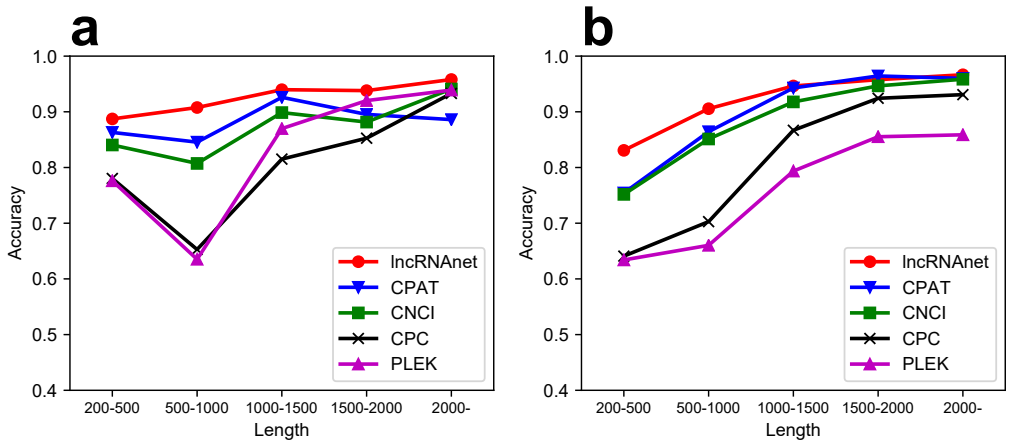

Figure 4.2 Accuracy of tools according to length variation in (a) HT and (b) MT

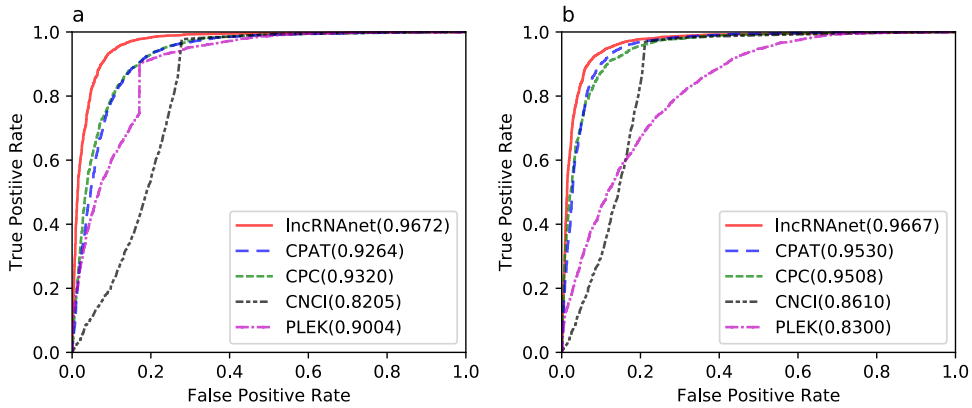

Figure 4.3 ROC comparison between tools and its AUC score
(a) ROC and AUC comparison in HT (b) ROC and AUC comparison in H2 (c) ROC and AUC comparison in MT

Additionally, we analyzed the effects of sequence length ranges. Figure 4.2(a) shows that the performance of lncRNAnet was the most accurate in all length ranges and consistently improved as the length of the sequences increased. On the other hand, the accuracy of other tools fluctuated according to the length variations.

4.3.2 Performance Comparison in a Cross-species Dataset

The dataset MT was used to evaluate the cross-species performance. The models of lncRNAnet and other tools except CPAT were trained on human data, while the model of CPAT was trained on mouse logistic regression.

Table 4.5 and Figure 4.3 (b) show that lncRNAnet still showed the best performance, 9.14%, 3.16%, 2.54%, and 1.44% higher than the second-best results in terms of specificity, accuracy, F1-score, and AUC, respectively. These results indicate that the model of lncRNAnet built from human data performed reasonably in a cross-species test. PLEK showed the lowest performance in all metrics.

In the length variation analysis, Figure 4.2(b) shows that the accuracy of all tools increased steadily as the sequence length increased. However, in terms of overall prediction accuracy, lncRNAnet showed the best performance.

The proposed lncRNAnet showed the best performance in terms of specificity, accuracy, F1-score, and AUC score in both independent test sets, HT, and MT. Furthermore, lncRNAnet performed robustly regardless of the length variations.

Chapter 5

Discussion

Deep learning-based approaches have been successfully applied in the bioinformatics domain [47]. RNN-applied tools based on shorter ncRNAs such as deepTarget [48] and deepMiRGene [49] outperformed existing tools. Although the length of the lncRNA transcripts was relatively longer, lncRNAnet also outperformed existing tools.

In the study, there were issues to consider while identifying lncRNAs. One was a sensitivity-specificity trade-off. In a binary classification, it is important to reach both high sensitivity and high specificity. Currently, the number of identified lncRNA transcripts is fairly small compared to the number of protein-coding transcripts. It is important to disregard false lncRNAs, but CPAT, CNCI, CPC, and PLEK had low specificity. They focus more on detecting lncRNAs and misclassify protein-coding transcripts, which require additional steps to filter protein-coding transcripts from predicted lncRNAs. In contrast, lncRNAnet with a balanced and high performance in terms of sensitivity and specificity can help reduce the additional filtering process.

Another issue was the accuracy change depending on the sequence length. When the sequence length increased, the accuracy of the tools, especially CPC and PLEK,

increased dramatically. The reason the performance of the other tools increased as the length increased is that the ORF features are dependent on the transcript length. An ORF of a noncoding transcript can be interpreted as a random sequence that is irrelevant to the function of coding. Briefly, a start codon and a stop codon may appear on average every 64 codons. Thus, the length of the ORF has a limit. As a result, the high performance on longer sequences benefits from the length difference between the ORF length and the transcript length. However, a large proportion of lncRNAs are shorter than protein-coding mRNAs, and the performance with shorter lncRNAs must be guaranteed. The proposed lncRNAnet robustly predicted lncRNAs on both short sequences and long sequences, which will contribute to discovering novel lncRNAs.

This study showed that RNNs perform successfully compared to the other approaches. RNNs can be improved by newly developed architectures, such as stack-augmented recurrent nets [50] and ByteNet [51]. Despite the outstanding performance of deep learning, it has weaknesses. Compared to the traditional machine learning approaches, a neural network is like a black box, which can identify only its input and output. The trained weights in the model can be identified, while the meaning of the features is difficult to interpret. There have been many efforts to analyze deep learning features. If the feature analysis is completed, we can discover new characteristics of lncRNAs.

Chapter 6

Conclusion

In this study, we proposed an RNN-based method for classifying lncRNAs from protein-coding transcripts. We used a novel feature, an ORF indicator, which is identified by one-dimensional CNNs, to reflect biological knowledge in our model. The proposed lncRNAnet showed 7.83%, 5.76%, 5.30%, and 3.78% improvements over the alternatives in terms of specificity, accuracy, F1-score, and AUC, respectively, on the test set HT. Furthermore, lncRNAnet successfully detected shorter lncRNAs and showed robust performance regardless of sequence length variations. Our method will contribute to the identification of novel lncRNAs from the abundant transcriptome data.

Bibliography

[1] John S Mattick. Non-coding rnas: the architects of eukaryotic complexity. *EMBO reports*, 2(11):986–991, 2001.

[2] John S Mattick and Igor V Makunin. Non-coding rna. *Human molecular genetics*, 15(suppl 1):R17–R29, 2006.

[3] Jeannie T Lee. Epigenetic regulation by long noncoding rnas. *Science*, 338(6113): 1435–1439, 2012.

[4] Roger P Alexander, Gang Fang, Joel Rozowsky, Michael Snyder, and Mark B Gerstein. Annotating non-coding regions of the genome. *Nature Reviews Genetics*, 11(8):559–571, 2010.

[5] Ken C Pang, Martin C Frith, and John S Mattick. Rapid evolution of noncoding rnas: lack of conservation does not mean lack of function. *Trends in Genetics*, 22 (1):1–5, 2006.

[6] Kevin Struhl. Transcriptional noise and the fidelity of initiation by rna polymerase ii. *Nature structural & molecular biology*, 14(2):103–105, 2007.

[7] Orly Wapinski and Howard Y Chang. Long noncoding rnas and human disease. *Trends in cell biology*, 21(6):354–361, 2011.

[8] Arunoday Bhan and Subhrangsu S Mandal. Long noncoding rnas: emerging stars in gene regulation, epigenetics and human disease. *ChemMedChem*, 9(9): 1932–1956, 2014.

[9] Chris P Ponting, Peter L Oliver, and Wolf Reik. Evolution and functions of long noncoding rnas. *Cell*, 136(4):629–641, 2009.

[10] Philipp Kapranov and Georges St Laurent. Dark matter rna: existence, function, and controversy. *Genomic "Dark Matter": Implications for Understanding Human Disease Mechanisms, Diagnostics, and Cures*, pages 7–15, 2012.

[11] Johnny TY Kung, David Colognori, and Jeannie T Lee. Long noncoding rnas: past, present, and future. *Genetics*, 193(3):651–669, 2013.

[12] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, et al. The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–1789, 2012.

[13] Jinfeng Liu, Julian Gough, and Burkhard Rost. Distinguishing protein-coding from non-coding rnas through support vector machines. *PLoS Genet*, 2(4):e29, 2006.

[14] 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.

[15] 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.

[16] Liang Sun, Haitao Luo, Dechao Bu, Guoguang Zhao, Kuntao Yu, Changhai Zhang, Yuanning Liu, Runsheng Chen, and Yi Zhao. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research*, page gkt646, 2013.

[17] Kun Sun, Xiaona Chen, Peiyong Jiang, Xiaofeng Song, Huating Wang, and Hao Sun. iseerna: identification of long intergenic non-coding rna transcripts from transcriptome sequencing data. *BMC genomics*, 14(2):1, 2013.

[18] 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, 2014.

[19] Supatcha Lertampaiporn, Chinae Thammarongtham, Chakarida Nukoolkit, Boonserm Kaewkamnerdpong, and Marasri Ruengjitchatchawalya. Identification of non-coding rnas with a new composite feature in the hybrid random forest ensemble algorithm. *Nucleic acids research*, page gku325, 2014.

[20] Rujira Achawanantakun, Jiao Chen, Yanni Sun, and Yuan Zhang. Lncrna-id: Long non-coding rna identification using balanced random forests. *Bioinformatics*, 31 (24):3897–3905, 2015.

[21] Cong Pian, Guangle Zhang, Zhi Chen, Yuanyuan Chen, Jin Zhang, Tao Yang, and Liangyun Zhang. Lncrnapred: Classification of long non-coding rnas and protein-coding transcripts by the ensemble algorithm with a new hybrid feature. *PloS one*, 11(5):e0154567, 2016.

[22] 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.

[23] Rashmi Tripathi, Sunil Patel, Vandana Kumari, Pavan Chakraborty, and Pritish Kumar Varadwaj. Deeplnc, a long non-coding rna prediction tool using deep neural network. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5 (1):1–14, 2016.

[24] Mitchell Guttman, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, Or Zuk, Bryce W Carey, John P Cassady, et al. Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, 458(7235):223–227, 2009.

[25] Tim R Mercer, Marcel E Dinger, and John S Mattick. Long non-coding rnas: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.

[26] Kevin C Wang and Howard Y Chang. Molecular mechanisms of long noncoding rnas. *Molecular cell*, 43(6):904–914, 2011.

[27] Jeffrey J Quinn and Howard Y Chang. Unique features of long non-coding rna biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016.

[28] Jeremy E Wilusz, Hongjae Sunwoo, and David L Spector. Long noncoding rnas: functional surprises from the rna world. *Genes & development*, 23(13):1494–1504, 2009.

[29] Marcel E Dinger, Ken C Pang, Tim R Mercer, and John S Mattick. Differentiating protein-coding and noncoding rna: challenges and ambiguities. *PLoS Comput Biol*, 4(11):e1000176, 2008.

[30] Barbara Hrdlickova, Rodrigo Coutinho de Almeida, Zuzanna Borek, and Sebo Withoff. Genetic variation in the non-coding genome: Involvement of micro-rnas

and long non-coding rnas in disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1910–1922, 2014.

[31] Igor Ulitsky and David P Bartel. lincrnas: genomics, evolution, and mechanisms. *Cell*, 154(1):26–46, 2013.

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013.

[35] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[36] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[37] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[38] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[39] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[40] Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks. http://karpathy.github.io/2015/05/21/rnn-effectiveness/, 2015. Online; accessed 11-December-16.

[41] Viacheslav Khomenko, Oleg Shyshkov, Olga Radyvonenko, and Kostiantyn Bokhan. Accelerating recurrent neural network training using sequence bucketing and multi-gpu data parallelization. In *Data Stream Mining & Processing (DSMP), IEEE First International Conference on*, pages 100–103. IEEE, 2016.

[42] Pierre Baldi and Søren Brunak. Chapter 6. neural networks: applications. In *Bioinformatics: The Machine Learning Approach*. MIT press, 2001.

[43] 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.

[44] Christopher M Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[45] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[46] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in Bioinformatics*, page bbw068, 2016.

[48] Byunghan Lee, Junghwan Baek, Seunghyun Park, and Sungroh Yoon. deeptarget: End-to-end learning framework for microrna target prediction using deep recurrent neural networks. *arXiv preprint arXiv:1603.09123*, 2016.

[49] Seunghyun Park, Seonwoo Min, Hyunsoo Choi, and Sungroh Yoon. deepmirgene: Deep neural network based precursor microrna prediction. *arXiv preprint arXiv:1605.00017*, 2016.

[50] Armand Joulin and Tomas Mikolov. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems*, pages 190–198, 2015.

[51] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

국문 초록

순환신경망을 이용한 lncRNA 판별

Long noncoding RNA (lncRNA)는 생물학적 과정에서 중요한 조절 인자이다. LncRNA는 messenger RNA (mRNA)와 유사한 서열 특성을 공유하지만 전혀 다른 역할을 담당하여 생물 연구에 새로운 통찰을 제시하였다. Next-generation sequencing (NGS)의 발전으로 인해 lncRNA 전사체를 검출할 수 있게 되었으나 실험적으로 많은 전사체를 판별하기 위해 많은 시간과 비용을 요구하였고 이를 해결하기 위해 lncRNA 판별에 계산적 접근이 필요하였다.

lncRNA를 판별하기 위해, RNA 서열을 모델링하는 recurrent neural network (RNN)와 종결 코돈을 찾아 lncRNA를 판별하는 convolutional neural network (CNN)를 이용한 딥 러닝 기반의 lncRNAnet을 개발하였다. LncRNAnet은 다른 lncRNA 판별 기법에 비해 대부분의 전사체들이 분포되어 있는 짧은 길이에서 가장 좋은 성능을 보였다. 또한 lncRNAnet은 human test set (HT)에서도 특이도, 정확도, F1-score와 area under the curve (AUC)에서 최고 성능의 다른 기법을 각각 7.83%, 5.76%, 5.30%와 3.78%로 앞질렀다.

주요어: Long noncoding RNA (lncRNA), 순환신경망, 딥러닝

학번: 2015-20507