BMC Genomics

SOFTWARE

Open Access

PLEKv2: predicting lncRNAs and mRNAs based on intrinsic sequence features and the coding-net model

Aimin Li[1*†], Haotian Zhou[1†], Siqi Xiong[3*†], Junhuai Li[1], Saurav Mallik[2], Rong Fei[1], Yajun Liu[1], Hongfang Zhou[1], Xiaofan Wang[1], Xinhong Hei[1] and Lei Wang[1]

Abstract

Background Long non-coding RNAs (lncRNAs) are RNA transcripts of more than 200 nucleotides that do not encode canonical proteins. Their biological structure is similar to messenger RNAs (mRNAs). To distinguish between lncRNA and mRNA transcripts quickly and accurately, we upgraded the PLEK alignment-free tool to its next version, PLEKv2, and constructed models tailored for both animals and plants.

Results PLEKv2 can achieve 98.7% prediction accuracy for human datasets. Compared with classical tools and deep learning-based models, this is 8.1%, 3.7%, 16.6%, 1.4%, 4.9%, and 48.9% higher than CPC2, CNCI, Wen et al.'s CNN, LncADeep, PLEK, and NcResNet, respectively. The accuracy of PLEKv2 was > 90% for cross-species prediction. PLEKv2 is more effective and robust than CPC2, CNCI, LncADeep, PLEK, and NcResNet for primate datasets (including chimpanzees, macaques, and gorillas). Moreover, PLEKv2 is not only suitable for non-human primates that are closely related to humans, but can also predict the coding ability of RNA sequences in plants such as *Arabidopsis*.

Conclusions The experimental results illustrate that the model constructed by PLEKv2 can distinguish lncRNAs and mRNAs better than PLEK. The PLEKv2 software is freely available at https://sourceforge.net/projects/plek2/.

Keywords lncRNAs, Deep learning, PLEK, Coding-net

Background

Long non-coding RNAs (lncRNAs) are defined as transcripts with lengths of more than 200 nucleotides without any protein-coding ability [1], and are poorly conserved between species, while messenger RNAs (mRNAs) are a category of transcripts that have similar sequence structure to lncRNAs but encode proteins. Whole transcriptome sequencing data suggests that more than 75% of the human genome and more than 50% of the *Arabidopsis* genome can be transcribed into RNAs [2, 3]. Non-coding RNAs (ncRNAs) account for more than 98% of the human genome and play important roles in gene expression and regulation [4].

†Aimin Li, Haotian Zhou and Siqi Xiong contributed equally to this work.

*Correspondence:
Aimin Li
liaiminmail@gmail.com
Siqi Xiong
1315459270@qq.com
[1]Shaanxi Key Laboratory for Network Computing and Security Technology, School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, Shaanxi 710048, China
[2]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]Department of Information Engineering, College of Technology, Hubei Engineering University, Xiaogan, Hubei 432000, China

Recent studies have shown that lncRNAs can play roles in the development of cancers. lncRNAs interact with DNA, other RNAs, and protein molecules and act as important regulatory factors at the epigenetic, transcriptional, and post-transcriptional levels [5]. Aberrant lncRNA expression can affect the initiation, growth, and metastasis of cancer cell tumors. For example, the lncRNA HOTAIR (HOX transcript antisense RNA) can promote the proliferation, survival, invasion, metastasis, and drug resistance of lung cancer cells [6].

As sequencing technology develops, many novel animal and plant transcripts have been sequenced [7]. The rapid and effective identification of lncRNAs among these transcripts is the basis for subsequent gene function analysis and evolution, requiring an efficient tool for identifying lncRNAs in animals and plants [8]. The performance of the alignment-free tool we developed previously, PLEK, could be improved in some species. For example, the accuracy of PLEK for mouse datasets is <90% [9], while the accuracy of PLEK for the *Arabidopsis* protein-coding dataset is <65%. Therefore, in this study, to solve these shortcomings, we upgrade PLEK to PLEKv2, and constructed models tailored for both animals and plants.[1] Here, we develop a novel deep learning classification model, called the "Coding-Net model", which fuses features of the calibrated open reading frame (ORF) lengths and multiple *k*-mer frequencies, and then use this Coding-Net classification model to identify lncRNAs and mRNAs.

Compared with traditional machine learning models, deep learning classification models do not require manual intervention. The intrinsic characteristics of the sequences can be learned independently and useful information extracted to improve the accuracy of prediction. The latest deep learning-based methods include the deep neural network (DNN) [10], convolutional neural network (CNN) [11], recurrent neural network (RNN) [12], deep belief network (DBN) [13], and residual neural network (ResNet) [14]. Deep learning-based classification tools include lncRNAnet [15], which uses one-dimensional CNNs to detect ORFs that are candidates for coding transcripts and an RNN as the classifier. Fan et al. proposed the LncRNA_Mdeep model [16] to predict lncRNAs by fusing a CNN and DNNs. Although these two deep learning-based methods perform better than previous conventional machine learning algorithms, they still depend on one-hot encoding for feature extraction, which will lead to sparse coding and a high computational cost. LncADeep, a recently developed deep learning-based method, can predict lncRNAs by fusing multiple features with a DBN as the classifier [17].

Compared with PLEK, PLEKv2 gives a high accuracy rate for both human (PLEK, 93.8%; PLEKv2, 98.7%) and mouse (PLEK, 88.3%; PLEKv2, 94.6%) datasets. This is partly because PLEKv2 includes the calibrated ORF lengths and uses the Coding-Net model for classification. PLEKv2 is particularly suitable for evaluating the coding ability of primate and plant transcriptomes. For additional validation, PLEKv2 was compared with the classical tools, CPC2 [18] and CNCI [19], and the deep learning-based models, Wen et al.'s CNN [20], LncADeep [17], and NcResNet, for human datasets. We found that PLEKv2 shows the highest accuracy of all the tools, with an accuracy of 98.7%. The open-source code of PLEKv2 is available online at https://sourceforge.net/projects/plek2/.

Implementation

Data description

We downloaded human lncRNA (Release 38) and mRNA (Release 206) transcript data from the GENCODE [21] and RefSeq [22] databases to construct human models. From these two datasets, equal numbers of transcripts were divided randomly into the training, validation, and test sets at a ratio of 8:1:1. The validation set was used to check the quality of the model and constantly upgrade the parameters of the iterative model. In addition, we also constructed a plant model using lncRNA and mRNA sequences downloaded from the Ensembl Plants database (v101) [23].

To evaluate the performance of PLEKv2 across species, we further constructed independent testing sets for both vertebrates and plants. The protein-coding transcripts were obtained from the RefSeq and Ensembl Plants databases, respectively, and sequences with 'putative', 'predicted', or 'pseudogene' annotations were excluded. Non-coding transcripts were obtained from the Ensembl [24] and Ensembl Plants databases.

Data pre-processing

RefSeq and GENCODE are widely used to provide biologically non-redundant and well-annotated sets of sequences. The two databases have different data collection and annotation strategies. The RefSeq database is considered the authority for annotating mRNA, while the GENCODE database is highly esteemed for lncRNA annotation because of its expertise and accuracy. Using both databases in combination provides a more comprehensive coverage of transcript information to build high-quality training and test datasets. The data pre-processing is described here in brief and shown in Fig. 1.

First, we removed sequences of less than 200 nucleotides from the original files, obtaining 48,471 human lncRNAs and 60,246 mRNAs. Then, we replaced all occurrences of 'U' with 'T' and replaced all the various mixed-base symbols with the mixed-base N. Mixed-base symbols indicate bases that are not completely determined, such as the R, Y, M, K, S, W, H, B, V, D, and N

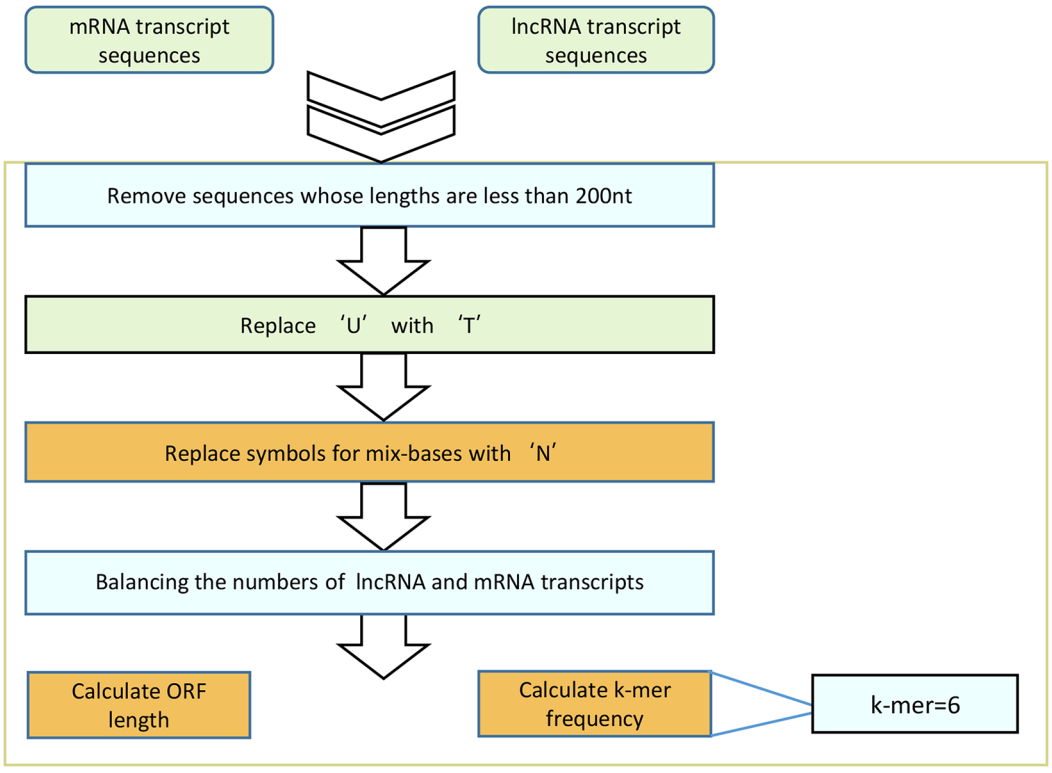

Fig. 1 Data pre-processing flowchart

bases. Then, to balance the transcript numbers, we randomly sampled the larger mRNA sample set to equalize the transcript numbers of lncRNAs and mRNAs, selecting 48,471 mRNA transcript sequences randomly. This step is key, as if the number of samples in one category differs significantly from other categories, the model may be more inclined to the dominant category. Finally, the weighted *k*-mer feature and ORF lengths were computed for the final set of sequences.

Open reading frames

ORFs contain a start codon (ATG) and one of the stop codons, and indicate the potential of a transcript sequence to encode protein. The ORF is therefore one of the most important basic characteristics for identifying protein-coding sequences, and is the main criterion for predicting the potential coding ability [25]. Prior studies have shown that lncRNAs have little or no ORF [26]. Thus, we added the ORF lengths to the feature vectors.

We used regular expressions to search for start codons (ATG) within the transcript to determine the ORF starting position. Then, starting from this position, we translated the sequence until we encountered a stop codon. If a complete ORF was found, including both start and stop codons, we calculated the length of this ORF, *l*, and processed and normalized it to balance the weight of the ORFs and *k*-mers. The process of calculating the ORF lengths was as follows. We obtained l_i, the logarithm (base 10) of the length of each peptide chain (l), and then normalized this with the min–max method to give the corrected ORF length value, X_{ORF}. The formulae were as follows:

$$l_i = \lg l \tag{1}$$

$$X_{ORF} = \frac{l_i - l_{\min}}{l_{\max} - l_{\min}} \tag{2}$$

where l_{min} is the minimum value of the peptide chain after taking the logarithm, and l_{max} is the maximum value of the peptide chain after taking the logarithm.

k-mers

lncRNAs are poorly conserved and have poor coding ability [27]. Our previous research has shown that the *k*-mer frequency may distinguish lncRNAs from mRNAs. A *k*-mer pattern is a specific string of *k* nucleotides, where each sequence can contain A, T, G, and C. When $k=1$ to 6, there are $4+16+64+256+1024+4096=5460$ different *k*-mer patterns: four monomer patterns, 16 two-mer patterns, 64 three-mer patterns, 256 four-mer patterns, 1024 five-mer patterns, and 4096 six-mer patterns. Figure 2 shows the sliding window process whereby we moved a sliding window along the transcript. For example, when

Fig. 2 The 6-mer sliding window showing the process of taking a *k*-mer in sliding window mode from a sequence when $k=6$

k is three, the length of the window is three and the step length is one.

Here, we used the feature vector X_{k-mer}, defined below, to represent the *k*-mer ($k=1, 2, 3, 4, 5, 6$) weighted frequencies in the transcript sequences, *f*. The calculation process was the same as in our previous study [9].

$$X_{k-\text{ mer}} = [f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}, f_{11} \cdots f_{5460}] \quad (3)$$

Construction of feature vectors

Feature fusion can obtain the most discriminatory information from an original multi-feature set and eliminate redundant information caused by correlations between different feature sets, improving the final outcome. Here, two types of feature sets for the ORFs and *k*-mers were concatenated into one set of feature vectors to represent the transcript sequences, as shown by the following formula:

$$X = [X_{k-\text{mer}}, X_{\text{ORF}}] \quad (4)$$

Classification using coding-net architecture

In this study, the Coding-Net model for predicting the transcript coding ability based on feature vectors was constructed from the *k*-mer frequencies and ORF lengths using convolutional neural network algorithms. The network structure of the model is shown in Fig. 3. The convolution layer is used for local perception, which uses a convolution kernel for feature extraction and feature mapping. The pooling layer is used for down sampling, sparse processing of feature maps, and reducing the amount of data computation. Concurrently, the dense layer is used to classify sequences.

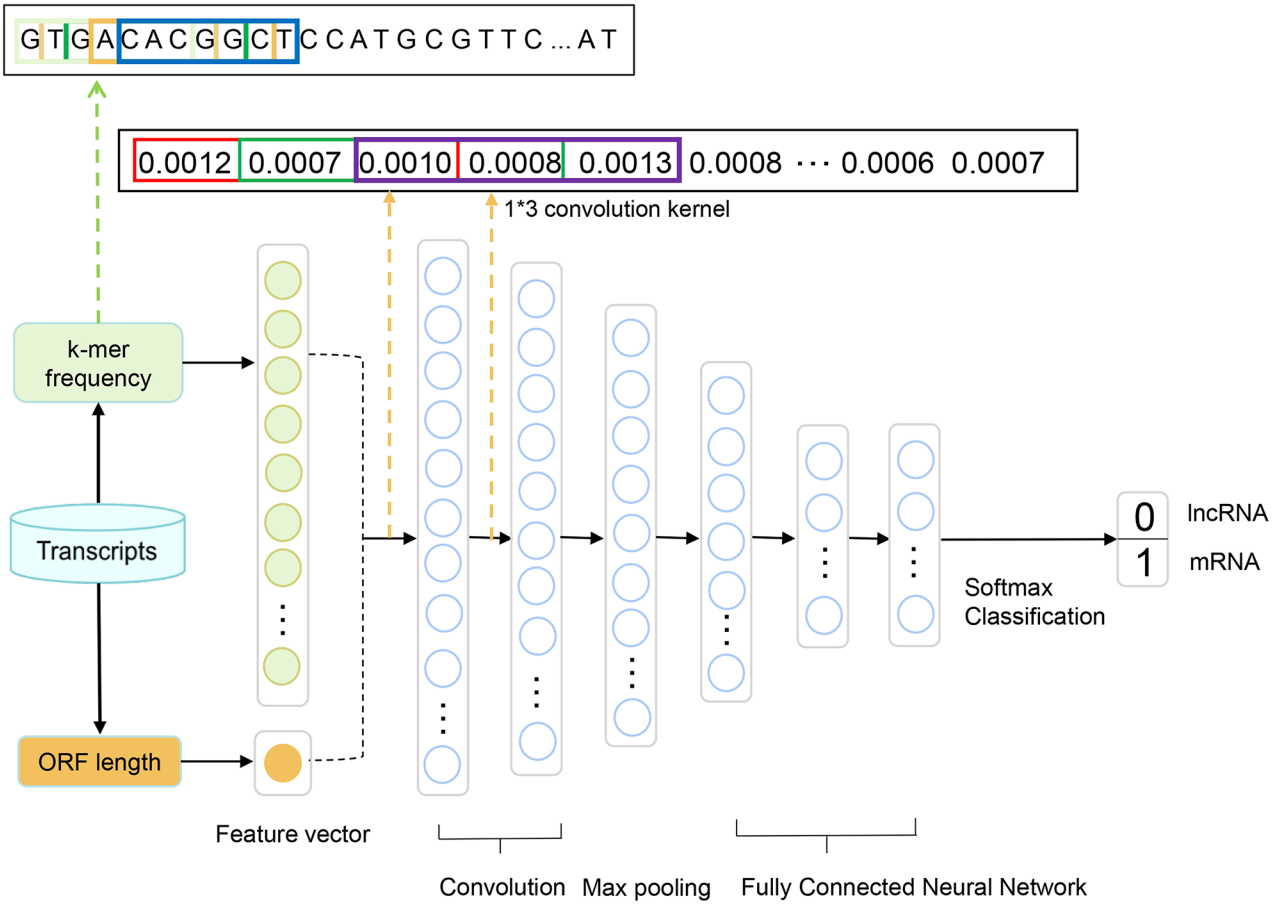

Fig. 3 The network structure of the Coding-Net model, which includes the input and deep neural networks. The input contains the *k*-mer frequency and calibrated ORF length derived from the sequence, and the deep neural network consists of three parts, including the convolution, max pooling, and dense layers

Table 1 The range of each hyper-parameter

Hyper-parameters	Range
Kernel size for convolution	1*3
Number of kernels	32
Pooling method	Max pooling
Number of units in hidden layer	265, 64 ,64
Optimizer	Adam

As shown in Fig. 3, the Coding-Net classification model includes the input and deep neural networks. The input contains the *k*-mer frequency and calibrated ORF length derived from the sequence, which are constructed into a feature vector. In previous studie, to apply convolutional neural networks to sequence data, one-dimensional sequence vectors have been randomly shuffled into high-dimensional matrices [20]. This approach may destroy the positional information of the sequence, and is not conducive to the extraction of classification features. To solve this problem, the Coding-Net model uses one-dimensional feature vectors as inputs with no reshaping of the dimension.

The deep neural network consists of three parts. First, it consists of two convolutional neural network layers. In the *k*-mer and ORF feature extraction layers, the input of each neuron was connected to the local region of the previous layer, and the local features were extracted using a 1×3 receptive field. The receptive field of each feature map was a plane, and all neurons used the same weight value. Second, maximum pooling was used to retain the main features while reducing the number of parameters in the network. This can also effectively reduce the occurrence of overfitting and the cost of computing resources. In addition, both the convolution and pooling layers were mapped to a hidden layer feature space by feature vectors. Finally, the three fully connected layers were used to classify lncRNAs and mRNAs, combine all local features into a complete feature map, and map the learned distributed features into the sample marker space.

During training of Coding-Net, five hyper-parameters (the convolution kernel size, the number of kernels, the pooling method, the number of units in the hidden layer, and the learning algorithm) were tuned. Table 1 lists the specific settings for each hyper-parameter. Hyper-parameters cannot be learned from the data in the standard model training process directly and need to be defined in advance.

This model was implemented using Keras [28]. We applied ReLU as the activation function. In all cases, we used a batch size of 128, and the cross-entropy loss was optimized using Adam.

Feature extraction by asymmetric convolution kernel

The asymmetric convolution kernel is a new structure to replace the standard convolution layer widely used in modern convolutional neural networks, the 3×3 convolution kernel. Specifically, a $1 \times d$ kernel is used to replace the $d \times d$ kernel [29]. This can greatly reduce the amount of computation required and enhance the robustness of the model, allowing an increased amount of significant information to be extracted.

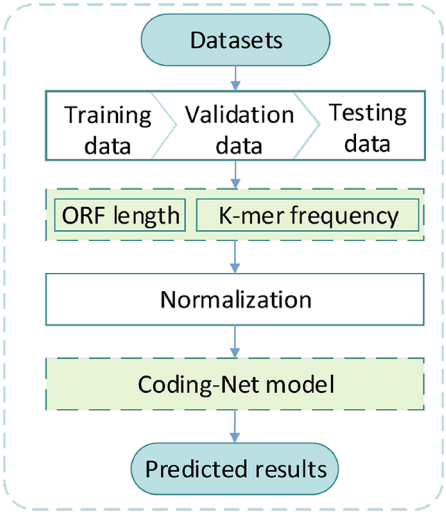

Fig. 4 The workflow of PLEKv2 tools

In this study, a 1×3 asymmetric convolution kernel was used to enhance the key differential features of the feature vectors and weaken the influence of irrelevant features. The 1×3 convolution kernel can be regarded as a codon, which is composed of three adjacent nucleotides in mRNA, and can more effectively extract and identify lncRNA features from the sequences. The workflow of the PLEKv2 tool is shown in Fig. 4.

Evaluation metrics.

To evaluate the predictive performance of PLEKv2, we used four evaluation indicators: precision, recall, the F_1 score, and accuracy. To highlight the importance and characteristics of ncRNA, for a better understanding of its role in biological processes, we defined non-coding as 'positive' and protein-coding proteins as 'negative'. The precision measures the ratio of true positives to all predicted positives, the recall measures the ratio of true positives that are correctly identified, and the F_1 score is a composite measure used as an aggregated performance score for the evaluation of algorithms.

Results

Optimal feature vectors for the coding-net classification model

Considering different *k*-mers as features will affect the accuracy of identification of lncRNAs and mRNAs. For the human dataset, we constructed a feature vector with only the weighted *k*-mer frequencies and fed it into the deep learning classification model. When *k* takes a

different value (k=5–6), the accuracy will increase as k increases. When k=5, the model achieved an accuracy of 95.4% with a size of 1×1364, and took 1.5 h to train. When k=6, the accuracy increased to 96.7%, with a larger size of 1×5460 (Table 2), and the training time extended to 5.2 h. If k=7, there would be 16,384 different k-mers, four times the number of 6-mers. This not only implies a significant increase in computational resource requirements but also indicates an increase in model complexity.

The ORF length is an effective differential feature for distinguishing between lncRNAs and mRNAs. To further understand the influence of the ORF on the internal decision of the model, the ORF alone gave an accuracy of 93.8% (Table 2). However, by fusing the k-mer (k=6) with the ORF features, the size of the feature vector was 1×5461. Compared with the unfused feature vectors, the accuracy rate increased from 96.7 to 98.5% for human datasets.

In conclusion, we used the k-mer (k=6) and normalized ORF length to construct feature vectors to train the Coding-Net model. Here, the size of the feature vector was 1×5461.

Table 2 The influence of feature vectors of different sizes on model accuracy

k value	Number of k-mers	ORF	Matrix form	Model accuracy
0	0	Yes	1*1	93.8%
5	1364	No	1*1364	95.4%
5	1364	Yes	1*1365	97.2%
6	5640	No	1*5460	96.7%
6	5640	Yes	1*5641	98.7%

Table 3 Performance of CPC2, CNCI, Wen et al.'s CNN, LncADeep, PLEK, NcResNet and PLEKv2 for human datasets

Models	Precision	Recall	F_1score	Accuracy
CPC2	0.942	0.856	0.897	0.906
CNCI	0.914	0.975	0.944	0.950
CNN	0.792	0.821	0.806	0.821
LncADeep	0.960	0.980	0.970	0.973
PLEK	0.962	0.941	0.938	0.938
PLEKv2	**0.986**	**0.986**	**0.986**	**0.987**
NcResNet	0.492	0.498	0.496	0.498

Performance of PLEKv2

Having determined the accuracies of our PLEKv2 model, we compared PLEKv2 with other state-of-the-art tools and models using human datasets (Table 3), including the ab initio tools CPC2, CNCI, and PLEK, and deep learning models such as Wen et al.'s CNN, LncADeep, and NcResNet.

As shown in Table 3, compared with other tools on the same test sets, PLEKv2 had the highest F_1 score (while the F_1 scores for CPC2, Wen et al.'s CNN, and NcResNet were all lower than 0.9), the highest precision of 0.986 (with the precision of Wen et al.'s CNN and NcResNet much lower), and the highest recall of 0.986 (with the recall of CPC2 and NcResNet much lower). Our method also showed the highest accuracy of these lncRNA identification tools, and clearly outperforms existing tools.

These results show that PLEKv2 performs better on human data than PLEK and other tools.

Cross-species prediction

At present, the genome sequences of most organisms are incomplete, and annotations are poor quality or unavailable. To solve this problem, well-annotated species can be used to make cross-species predictions for other non-model organisms. In this study, we used human data to analyze data from other organisms that have not been explored in depth.

We used eight organisms to test the cross-species predictive performance of CPC2, CNCI, PLEK, PLEKv2, LncADeep, and NcResNet (Table 4). The results show that the overall performance of PLEKv2 was better than that of PLEK. PLEKv2 performed best on three species. While the accuracies of PLEKv2 and LncADeep were >90%, LncADeep demonstrated a higher average accuracy, but with a run time 11 times longer than PLEKv2. The accuracy of NcResNet was <80% on all datasets. The accuracy of CPC2 was 85.5% for *Pan troglodytes*, while the accuracy of CNCI was 89.4% for *Rattus norvegicus*. These examples clearly demonstrate the poor predictive performance of CPC2 and CNCI for specific species. These results indicate that PLEKv2 and LncADeep exhibit good performance for cross-species prediction.

Table 4 The accuracy of CPC2, CNCI, PLEK, PLEKv2, LncADeep, and NcResNet for cross-species prediction

Species	Number of transcripts	CPC2	CNCI	LncADeep	PLEK	PLEKv2	NcResNet
Bos taurus	64,906	92.3%	93.6%	**96.8%**	91.3%	93.8%	52.7%
Gorilla gorilla	33,667	91.8%	87.4%	90.5%	83.8%	**92.2%**	52.5%
Macaca mulatta	12,006	92.6%	94.5%	93.2%	87.3%	**95.2%**	50.3%
Mus musculus	41,588	94.1%	95.1%	**97.1%**	88.3%	94.6%	70.0%
Pan troglodytes	4,506	87.9%	91.3%	93.4%	90.4%	**93.5%**	51.1%
Rattus norvegicus	20,903	91.3%	89.4%	**95.5%**	88.8%	91.3%	51.3%
Sus scrofa	13,379	93.4%	94.6%	**95.6%**	75.7%	91.5%	50.9%
Xenopus tropicalis	8,669	**98.5%**	96.6%	98.7%	96.8%	97.3%	51.4%

Table 5 Performance comparison for primate datasets, with the best performances (precision, recall, F_1 score, and accuracy) between CPC2, CNCI, PLEK, LncADeep, NcResNet, and PLEKv2 shown in bold

Species	Tool	Precision	Recall	F_1score	Accuracy
Pan troglodytes	CPC2	0.755	0.938	0.837	0.879
	CNCI	0.849	0.899	0.873	0.913
	LncADeep	0.870	0.939	0.903	0.934
	PLEK	0.842	0.872	0.856	0.904
	PLEKv2	**0.873**	**0.940**	**0.905**	**0.935**
	NcResNet	0.343	0.532	0.417	0.511
Macaca mulatta	CPC2	0.954	0.902	0.927	0.926
	CNCI	0.937	**0.966**	0.951	0.945
	LncADeep	**0.968**	0.913	0.944	0.932
	PLEK	0.882	0.885	0.883	0.873
	PLEKv2	0.948	0.957	**0.952**	**0.952**
	NcResNet	0.544	0.489	0.516	0.503
Gorilla gorilla	CPC2	0.998	0.917	0.955	0.918
	CNCI	0.998	0.874	0.932	0.874
	LncADeep	**0.999**	0.905	0.950	0.905
	PLEK	**0.999**	0.838	0.911	0.838
	PLEKv2	**0.999**	**0.922**	**0.959**	**0.922**
	NcResNet	0.981	0.525	0.684	0.525

Performance comparison for primate datasets

We compared the performance of PLEKv2 with that of PLEK, CPC2, CNCI, LncADeep, and NcResNet for primates, including *Pan troglodytes* and *Macaca mulatta*. For chimpanzee (*Pan troglodytes*) datasets, 1485 protein-coding transcripts and 3021 non-coding transcripts were selected. For the macaque (*Macaca mulatta*) dataset, 5515 protein-coding transcripts and 6491 non-coding transcripts were selected from the RefSeq and Ensembl datasets. For the dataset for *Gorilla gorilla*, 33,025 protein-coding transcripts and 642 non-coding transcripts were selected from the RefSeq and Ensembl datasets.

On primate datasets, PLEKv2 achieved the highest accuracies (93.5% for chimpanzees, 95.2% for macaques, and 92.2% for gorillas), surpassing PLEK (90.4%, 87.3%, and 83.8%, respectively), CPC2 (87.9%, 92.6%, and 91.8%, respectively), CNCI (91.3%, 94.5%, and 87.4%, respectively), LncADeep (93.4%, 93.2%, and 90.5%, respectively), and NcResNet (51.1%, 50.3%, and 52.5%, respectively) (Table 5). Although LncADeep achieved the highest precision for the macaque dataset (96.8%) and CNCI achieved the highest recall for the macaque dataset (96.6%), the differences between PLEKv2, LncADeep, and CNCI were all within 2%. These results indicate that the PLEKv2 tool is particularly suitable for use in primates.

Performance comparison for mouse datasets

Mice are evolutionarily close to humans, and they also possess a relatively abundant supply of experimentally validated lncRNAs and mRNAs [17]. The accuracy of PLEK for identifying protein-coding transcripts and non-coding transcripts was <90% in mouse datasets, while PLEKv2 has a total accuracy of 94.6%.

To further evaluate the performance of PLEKv2 on mouse datasets, we compared PLEKv2 with PLEK, CPC2, CNCI, LncADeep, and NcResNet on the same datasets (consisting of 35,999 mRNAs and 5589 lncRNAs). Figure 5 shows the fraction of coding and non-coding transcripts marked by each tool, revealing that the accuracy of PLEKv2 was much higher than that of PLEK for mouse datasets.

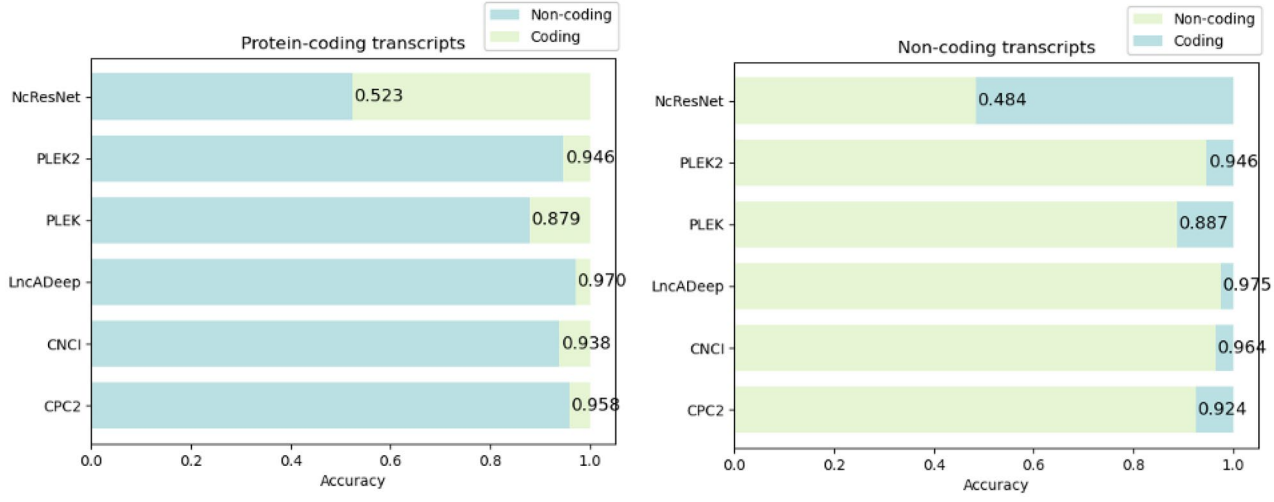

Fig. 5 Results of PLEKv2, PLEK, CPC2, and CNCI for mouse datasets. (**a**) The fraction of the protein-coding transcripts classified as coding or non-coding. (**b**) The fraction of the non-coding transcripts classified as coding or non-coding. This shows that the accuracy of PLEKv2 is much higher than that of PLEK for mouse datasets. CPC2 and PLEKv2 outperformed CNCI and PLEK for protein-coding transcripts, while PLEKv2 and CNCI outperformed CPC2 and PLEK for non-coding transcripts

Table 6 Accuracies of CPC2, PLEK, and PLEKv2 for plant datasets, with the highest accuracy between CPC2, PLEK, and PLEKv2 shown in bold

Species	Dataset type	Number of transcripts	CPC2	PLEK	PLEKv2
Arabidopsis thaliana	Coding	388	85.90%	60.2%	**95.7%**
	Non-coding	388	**97.30%**	91.20%	95.7%
Arabidopsis lyrata	Coding	37,026	94.20%	62.90%	**96.9%**
	Non-coding	795	95.60%	**100%**	98.2%
Oryza sativa	Coding	37,389	**96.50%**	78.90%	95.30%
	Non-coding	1011	100%	100%	100%

CPC2, PLEKv2, and LncADeep outperformed CNCI, PLEK, and NcResNet for protein-coding RNAs. PLEKv2, CNCI, and LncADeep outperformed CPC2, PLEK, and NcResNet for non-coding transcripts. LncADeep gave the highest accuracy for the mouse dataset.

It was of note that CNCI took two hours to predict the protein-coding transcripts of mouse, LncADeep took one hour to predict protein-coding transcripts in mice, and PLEKv2 only needed a few minutes for the mouse datasets. These results all indicate that PLEKv2 is a more efficient tool, able to identify lncRNAs and mRNAs more quickly.

Comparison with plant datasets

To make the application range of PLEKv2 wider, we used *Arabidopsis thaliana* to build a model for prediction in other plants that have not been explored in depth. We obtained *Arabidopsis thaliana* protein-coding and non-coding transcripts from the RefSeq and Ensembl Plants databases. Using the same pre-processing methods as for the human dataset, we obtained 3878 mRNAs and 3878 lncRNAs. These transcripts were randomly divided into training, validation, and test sets at a ratio of 8:1:1. We trained the Coding-Net model using the same hyper-parameter settings as for training the human dataset. The accuracy rate was 95.7%. To further explore the performance of PLEKv2, we also ran PLEK and CPC2, and the results are shown in Table 6. PLEKv2 gave an accuracy rate of >95.7% (average of coding and non-coding transcripts) on all plants, and showed a great improvement over PLEK in identifying protein-coding transcripts, from 62.9 to 96.9% in the *Arabidopsis lyrata* dataset, and 78.9–95.3% in *Oryza sativa*.

Although CPC2 performed better on both datasets, the difference between the accuracies of PLEKv2 and CPC2 was only about 2%, and on the *Arabidopsis* protein-coding dataset, PLEKv2 was 9.8% higher than CPC2. These results indicate that PLEKv2 maintains a high accuracy rate for predicting plant lncRNAs and mRNAs.

Predicting human RNAs containing short ORFs

Many molecules previously considered to be ncRNAs have been discovered to contain short peptides or small open reading frames (sORFs), which may play crucial roles in cellular functions and gene regulation [30]. Therefore, we used PLEKv2 to predict human RNAs containing short ORFs. The data we used came from CPPred [25], consisting of 641 coding RNAs and 641 lncRNAs. The results showed that PLEKv2 achieved a prediction accuracy of 89.2%, significantly higher than CPPred (accuracy=80.66%) [25]. This indicates that PLEKv2 maintains a high performance even when dealing with complex RNA structures.

Computational performance

We measured the computational performance of CPC2, CNCI, PLEK, PLEKv2, LncADeep, and NcResNet on a sample of 1000 protein-coding sequences and 1000 long non-coding sequences, randomly selected from the human RefSeq and GENCODE databases (Table 7), respectively. All the tools were run in a single-threading manner on an AMD Ryzen 7 4800U (16 cores @ 1.8 GHz) and 16 GB of RAM. PLEKv2 took 116 s to process the data, which was approximately 23 times faster than CNCI (2976 s), threefold faster than Wen et al.'s CNN (472 s), 11 times faster than LncADeep (1372 s), and four times faster than NcResNet (482 s). PLEKv2 also only requires 600 megabytes of memory to run.

Discussion

PLEKv2 uses two intrinsic sequence characteristics, the *k*-mer usage and calibrated ORF length, which are easily comprehensible and biologically meaningful. First, we constructed a feature vector with different weighted *k*-mer frequencies, and the accuracy of the Coding-Net model with just 6-mers was the highest, with an accuracy of 96.7% (Table 2). Next, we added the calibrated ORF length, and the accuracy rate increased from 96.7 to 98.7% for human datasets. These results indicate that the *k*-mer and ORF fused vectors are suitable classification features.

Table 7 Comparison of the computational performances of CPC2, CNCI, Wen et al.'s CNN, PLEK, PLEKv2, and LncADeep

Performance	CPC2	CNCI	CNN	PLEK	PLEKv2	LncADeep	NcResNet
Run time (seconds)	17	2976	472	128	116	1372	482
Memory (MB)	1080	24	819	300	660	90	2688
Online running	Yes	No	No	No	No	No	No

We used one-dimensional sequence vectors as inputs, preserving the positional information of sequences. Compared with the random shuffling of sequence features into high-dimensional matrices, one-dimensional input is more beneficial for the extraction of classification features. The range of the receptive field is related to the kernel size. The smaller the receptive field is, the more local and detailed its features tend to be. Here, we used 1×3 convolution kernels, which may have some biological significance and were regarded as codons, to enhance the differential feature information of neighboring nucleotides.

Compared with the established lncRNA and mRNA classification models, such as CPC2 and CNCI (machine learning algorithms), and Wen et al.'s CNN, LncADeep, and NcResNet (deep learning algorithms), the PLEKv2 model is excellent for human datasets. Meanwhile, PLEKv2 performs better than CPC2 at predicting ncRNAs in mice (*Mus musculus*), and also demonstrates superior performance for primate datasets.

There remain several limitations to our research. For example, other features aside from the ORF size, such as the ORF coverage and the ORF integrity, are not considered. Moreover, PLEKv2 exhibits relatively lower discriminative ability in determining the translational potential of transcripts. In addition, CPC2 exhibits more efficient performance than PLEKv2 in terms of run time. As an online web server, CPC2 is not only more user-friendly but can also be used as a stand-alone tool [18]. However, it is worth noting that CPC2 requires more memory resources during run time. Finally, The model was tested only on fully assembled and currently annotated datasets, without including more types of data such as incomplete transcripts.

Conclusions

Employing a novel classification model, we have upgraded our PLEK tool to its next version, PLEK version 2 (PLEKv2), based on deep learning algorithms. PLEKv2 is more accurate than PLEK, especially for primates and plants. More significantly, the accuracy of the PLEKv2 classification model was 98.7%, while CPC2 was 90.6%, CNCI was 95%, Wen et al.'s CNN was 82.1%, LncADeep was 97.3%, PLEK was 93.8%, and NcResNet was 49.8%. A very small improvement in accuracy is not a trivial matter: as there are very many lncRNAs, a 1% improvement in accuracy indicates the correct identification of hundreds more lncRNAs. With the development of third-generation sequencing technologies, increasing numbers of full transcripts are emerging, and PLEKv2 achieved high accuracy for lncRNA identification. Furthermore, PLEKv2 exhibits good performance in cross-species prediction, as do CPC2 and CNCI. The future scope of PLEKv2 will include further upgrades to higher versions, incorporating additional features and benefits that will greatly aid researchers worldwide.

Availability and requirements

Project name: PLEKv2.

Project home page: https://sourceforge.net/projects/plek2/.

Operating system: Linux/Unix.

Programming language: Python.

Other requirements: Python 3.8.5 or later versions.

install regex=2023.10.3 package.

install keras=2.4.3 package.

install pandas=2.0.3 package.

install tensorflow=2.4.1 package.

install bio>=1.3.2 package.

install numpy==1.19.2 package.

License: MIT.

Any restrictions to use by non-academics: None.

Abbreviations

lncRNA	Long non-coding RNA
mRNA	Messenger RNA
PLEK	Predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme
PLEKv2	Predicting lncRNAs and mRNAs based on intrinsic sequence features and the Coding-Net model
Coding-Net	A novel deep learning classification model
CPC2	Coding Potential Calculator 2
CNCI	Coding-Non-Coding Index
CNN	Convolutional Neural Network
LncADeep	Long non-coding RNA Annotation based on Deep learning (full-length)
NcResNet	Non-Coding Residual Neural Network
ORF	Open Reading Frame
DNN	Deep Neural Network
DBN	Deep Belief Network
RNN	Recurrent Neural Networks
ResNet	Residual Neural Network

Acknowledgements

Not applicable.

Author contributions

A.L. conceived and designed the study.A.L.,H.Z. and S. X. performed statistical analyses and wrote the paper. J.L.,R.F. and Y.L. participated in data analyses.S.M.,H.Z.,X.W.,X.H. and L.W. participated in the design of the study and contributed to acquisition of data. All authors read and approved the final manuscript.

Funding

This study was supported by Natural Science Basic Research Program of Shaanxi (2024JC-YBMS-484), the National Natural Science Foundation of China (61971347, 62202374, U21A20524, 62176146), and the National Natural Science Foundation International cooperation and exchange projects (62120106011).

Data availability

The open-source code of PLEKv2 is online available at https://sourceforge.net/projects/plek2/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 2 April 2024 / Accepted: 25 July 2024
Published online: 02 August 2024

References

1. Cuevas-Diaz Duran R, Wei H, Kim DH, Wu JQ. Invited review: Long non-coding RNA s: important regulators in the development, function and disorders of the central nervous system. Neuropathol Appl Neurobiol. 2019;45(6):538–56.
2. Wu L, Liu S, Qi H, Cai H, Xu M. Research progress on plant long non-coding RNA. Plants. 2020;9(4):408.
3. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F. Landscape of transcription in human cells. Nature. 2012;489(7414):101–8.
4. Berg JM, Tymoczko JL, Stryer L, Clarke ND. Biochemistry. Volume 5. WH freeman New York; 2002.
5. Liu SJ, Dang HX, Lim DA, Feng FY, Maher CA. Long noncoding RNAs in cancer metastasis. Nat Rev Cancer. 2021;21(7):446–60.
6. Loewen G, Jayawickramarajah J, Zhuo Y, Shan B. Functions of lncRNA HOTAIR in lung cancer. J Hematol Oncol. 2014;7:1–10.
7. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;12(2):87–98.
8. Li J, Zhang X, Liu C. The computational approaches of lncRNA identification based on coding potential: status quo and challenges. Comput Struct Biotechnol J. 2020;18:3666–77.
9. Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics. 2014;15(1):311.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
11. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012, 25.
12. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
13. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–54.
14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*: 2016; 2016: 770–778.
15. Baek J, Lee B, Kwon S, Yoon S. LncRNAnet: long non-coding RNA identification using deep learning. Bioinformatics. 2018;34(22):3889–97.
16. Fan XN, Zhang SW, Zhang SY, Ni JJ. lncRNA_Mdeep: an alignment-free predictor for distinguishing long non-coding RNAs from protein-coding transcripts by Multimodal Deep Learning. Int J Mol Sci 2020, 21(15).
17. Yang C, Yang L, Zhou M, Xie H, Zhang C, Wang MD, Zhu H. LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. Bioinformatics. 2018;34(22):3825–34.
18. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, Gao G. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45(W1):W12–6.
19. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res. 2013;41(17):e166–166.
20. Wen J, Liu Y, Shi Y, Huang H, Deng B, Xiao X. A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. BMC Bioinformatics. 2019;20(1):1–14.
21. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I. GENCODE 2021. Nucleic Acids Res. 2021;49(D1):D916–23.
22. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
23. Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, Bhai J, Campbell LI, Carbajo Martinez M, Chakiachvili M. Ensembl genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 2022;50(D1):D996–1003.
24. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):D884–91.
25. Tong X, Liu S. CPPred: coding potential prediction based on the global description of RNA sequence. Nucleic Acids Res. 2019;47(8):e43–43.
26. Sato Rgergpitcphnh-Ktkhnmsn, 13 Mgscwrlerjbe. 6 SmHY: analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002;420(6915):563–73.
27. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013;154(1):26–46.
28. Ketkar N, Santana E. Deep learning with Python. Volume 1. Springer; 2017.
29. Ding X, Guo Y, Ding G, Han J. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: *Proceedings of the IEEE/CVF international conference on computer vision: 2019*; 2019: 1911–1920.
30. Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367(6482):1140–6.

Publisher's Note