lncRNA_Mdeep: an alignment-free predictor for long non-coding RNAs identification by multimodal deep learning

Xiao-Nan Fan
Northwestern Polytechnical University
Shao-Wu Zhang (zhangsw@nwpu.edu.cn)
https://orcid.org/0000-0003-1305-7447
Song-Yao Zhang
Northwestern Polytechnical University
Jin-Jie Ni
Northwestern Polytechnical University

Research article

Keywords: Long noncoding RNA, alignment-free, multimodal learning, deep learning

Posted Date: May 13th, 2020

DOI: https://doi.org/10.21203/rs.2.16792/v2

Abstract

Background: Long non-coding RNAs (lncRNAs) play crucial roles in diverse biological processes and human complex diseases. Distinguishing lncRNAs from protein-coding transcripts is a fundamental step for analyzing lncRNA functional mechanism. However, the experimental identification of lncRNAs is expensive and time-consuming.

Results: In this study, we present an alignment-free multimodal deep learning framework (namely lncRNA_Mdeep) to distinguish lncRNAs from protein-coding transcripts. LncRNA_Mdeep incorporates three different input modalities (i.e. OFH modality, k-mer modality, and sequence modality), then a multimodal deep learning framework is built for learning the high-level abstract representations and predicting the probability whether a transcript is lncRNA or not.

Conclusions: LncRNA_Mdeep achieves 98.73% prediction accuracy in 10-fold cross-validation test on human. Compared with other eight state-of-the-art methods, lncRNA_Mdeep shows 93.12% prediction accuracy independent test on human, which is 0.94%~15.41% higher than that of other eight methods. In addition, the results on 11 cross-species datasets show that lncRNA_Mdeep is a powerful predictor for identifying lncRNAs. The source code can be downloaded from https://github.com/NWPU-903PR/lncRNA_Mdeep.

Background

LncRNAs are defined as non-protein-coding transcripts with the length more than 200 nucleotides. Several studies reveal that more than 70% of the human genome are capable of being transcribed, whereas less than 2% of the genome can be translated into proteins [1]. LncRNAs make up the largest portion of the non-protein-coding transcripts [2-4] and show critical roles in cellular function, development, and diseases [5-7].

LncRNA identification is the fundamental step of lncRNA-related researches, which has drawn a lot of attention in recent years. Several computational methods are developed for distinguishing lncRNAs from protein-coding transcripts. Existing computation methods can mainly categorized into alignment-based methods [8-13] and alignment-free methods [14-21]. The alignment-based methods generally align the transcripts against comprehensive reference protein database to predict lncRNAs, for example, CPC [8] aligned transcripts against UniRef90 dataset [22] using BLSATX [23] tool; lncRNA-ID [11] and lncADeep [13] aligned the transcripts against Pfam dataset [24] using HMMER [25] tool. This kind of methods heavily rely on the quality of alignments, which will be influenced by the performance of multiple-sequence alignment tools and the quality of reference databases. Furthermore, the alignment process is extremely time-consuming [15, 21]. To avoid the drawback caused by alignment, the alignment-free methods are developed to distinguish lncRNAs from protein-coding transcripts. Without considering conservation features, CNCI [14] extracted five features (i.e., the length and S-score of MLCDS, length-percentage, score-distance and codon-bias) by profiling adjoining nucleotide triplets to represent the

transcript sequences. CPAT [15] calculated open reading frame size, open reading frame coverage, Fickett TESTCODE score and hexamer score. PLEK [16] proposed an improved k-mer feature. And these methods adopted different machine learning algorithms to build the classifiers for predicting lncRNAs. For example, CNCI and PLEK used support vector machine (SVM), and CPAT used logistic regression. Except these conventional machine learning algorithms, Deep learning, a branch of machine learning, has been applied for lncRNA identification. Such as, lncRNA-MFDL [17] was developed to identify lncRNAs by fusing multiple features and a deep stacking network, and Tripathi et.al. [18] proposed the DeepLNC method to identify lncRNAs by k-mer features and a deep neural network classifier. Although deep learning algorithms achieve a better performance than conventional machine learning algorithms, these two methods still depend on manually crafted features, and fail to learn intrinsic features automatically from raw transcript sequences. Recently, a deep learning-based method, lncRNAnet [20], was proposed to identify lncRNAs. LncRNAnet built a convolutional neural networks (CNNs) for detecting the open reading frame (ORF) indicator and a recurrent neural network (RNNs) for modeling RNA sequence to identify lncRNAs, which does not take into consideration at all the manually crafted features.

In this study, we proposed an alignment-free method, lncRNA_Mdeep, to identify lncRNAs by using multimodal deep learning. The novelties of lncRNA_Mdeep mainly include: 1) lncRNA_Mdeep successfully integrates manually crafted features and raw transcript sequences. 2) LncRNA_Mdeep effectively extracts high-level abstract representations from multiple deep learning models based on different raw input features. 3) lncRNA_Mdeep successfully distinguish lncRNAs from protein-coding transcripts in not only human dataset but also multiple cross-species datasets. To validate our lncRNA_Mdeep, we tested it on human dataset containing 46,000 transcripts in 10-fold cross validation (10CV) test and compared it with other seven model architectures. Furthermore, we compared lncRNA_Mdeep with other eight state-of-the-art methods on the human and 11 cross-species datasets in independent test. The results show that lncRNA_Mdeep can effectively distinguish lncRNAs from protein-coding transcripts.

Results

We developed an alignment-free multimodal deep learning framework (namely lncRNA_Mdeep) to distinguish lncRNAs from protein-coding transcripts (Figure 1, Methods). In statistical prediction, the jackknife test, *q*-fold cross-validation (CV) test, and independent dataset test are often used to examine the effectiveness of a predictor in practical application [26]. Of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [27]. However, for large scale database, the jackknife test needs to spend lots of time to generate the prediction results. To reduce the computational time and evaluate the generalization performance of a predictor, in this study, we adopted the 10-fold cross-validation (10CV) test and independent dataset test as done by most investigators [17, 28-30]. For 10CV test, the transcripts in the training set are randomly partitioned into 10 subsets with approximately equal size, and one of the 10 subsets is singled out in turn as test transcripts and the other 9 subsets are used as the training transcripts. This process is repeated for 10 iterations, each time setting aside a different test subset. The results from the 10 folds can then be

averaged to produce a single estimation [31, 32]. For independent dataset test, all transcripts in testing set are outside the training set.

To evaluate the performance of lncRNA_Mdeep, we first investigated the performance of lncRNA_Mdeep with different model architectures on human dataset in 10CV test, and shown the effect of different hyper-parameters in DNNs and CNN, then compared lncRNA_Mdeep with eight existing state-of-the-art methods (i.e., CNCI [14], CPAT [15], PLEK [16], lncRNA-MEDL [17], CPC2 [19], lncRNAnet [20], LncFinder^1 and LncFinder^2 [21]) on human and 11 cross-species datasets in independent test. LncFinder^1 means the LncFinder without secondary structure, and LncFinder^2 means LncFinder with secondary structure.

LncRNA_Mdeep is implemented in python 3 using keras 2.2.4 [33] with the backend of Tensorflow-gpu (1.9.0) [34]. All the experiments are implemented on an Ubuntu system with a NVIDIA TITAN V GV100.

Performance of lncRNA_Mdeep

Performance of different model architectures

We separately implemented the DNN model with OFH feature as input (namely OFH_DNN), DNN model with k-mer feature as input (namely k-mer_DNN), CNN model with one-hot encoding as input (namely One-hot_CNN), the combinations of these models (i.e., OFH_DNN + k-mer_DNN, k-mer_DNN + One-hot_CNN, and OFH_DNN + One-hot_CNN), and the decision fusion of three models on Human training dataset in 10CV test. The results are shown in Table 1, from which we can see that the accuracy, S_n, S_p and MCC of lncRNA_Mdeep are 98.73%, 98.95%, 98.52% and 0.9748, respectively. By comparing the performance of OFH_DNN, k-mer_DNN, One-hot_CNN and lncRNA_Mdeep, we found that the accuracy of lncRNA_Mdeep is 2.99%, 2.20%, and 2.91% higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. The MCC of lncRNA_Mdeep is 0.0577, 0.0441, and 0.0579 higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. The S_n of lncRNA_Mdeep is 4.51 %, 2.55%, and 1.94% higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. The S_p of lncRNA_Mdeep is 1.48%, 1.86%, and 3.89% higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. These results show that lncRNA_Mdeep through incorporating three different input modalities achieves better performance than individual models. In addition, k-mer_DNN shows the best performance among three individual models (i.e., OFH_DNN, k-mer_DNN, and One-hot_CNN).

By comparing the performance of different combination of three individual models and lncRNA_Mdeep, we found that the accuracy of lncRNA_Mdeep is 2.76%, 0.37%, and 1.13% higher than that of OFH_DNN + k-mer_DNN, k-mer_DNN + One-hot_CNN, and OFH_DNN + One-hot_CNN, respectively. The MCC of lncRNA_Mdeep is 0.0537, 0.0074, and 0.0222 higher than that of OFH_DNN + k-mer_DNN, k-mer_DNN + One-hot_CNN, and OFH_DNN + One-hot_CNN, respectively. These results show that lncRNA_Mdeep through fusing three models achieves better performance than that of fusing any two models.

Furthermore, we also compared lncRNA_Mdeep with a decision fusion strategy of voting. As shown in Table 1, the performance of lncRNA_Mdeep is 0.31% and 0.0059 higher than that of voting fusion

strategy in terms of accuracy and MCC. All the results from Table 1 show that lncRNA_Mdeep is a superior deep learning framework and it can effectively distinguish lncRNAs from protein-coding transcripts.

Performance of lncRNA_Mdeep

Performance of different model architectures

We separately implemented the DNN model with OFH feature as input (namely OFH_DNN), DNN model with k-mer feature as input (namely k-mer_DNN), CNN model with one-hot encoding as input (namely One-hot_CNN), the combinations of these models (i.e., OFH_DNN + k-mer_DNN, k-mer_DNN + One-hot_CNN, and OFH_DNN + One-hot_CNN), and the decision fusion of three models on Human training dataset in 10CV test. The results are shown in Table 1, from which we can see that the accuracy, S_n, S_p and MCC of lncRNA_Mdeep are 98.73%, 98.95%, 98.52% and 0.9748, respectively. By comparing the performance of OFH_DNN, k-mer_DNN, One-hot_CNN and lncRNA_Mdeep, we found that the accuracy of lncRNA_Mdeep is 2.99%, 2.20%, and 2.91% higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. The MCC of lncRNA_Mdeep is 0.0577, 0.0441, and 0.0579 higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. The S_n of lncRNA_Mdeep is 4.51 %, 2.55%, and 1.94% higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. The S_p of lncRNA_Mdeep is 1.48%, 1.86%, and 3.89% higher than that of OFH_DNN, k-mer_DNN, and One-hot_CNN, respectively. These results show that lncRNA_Mdeep through incorporating three different input modalities achieves better performance than individual models. In addition, k-mer_DNN shows the best performance among three individual models (i.e., OFH_DNN, k-mer_DNN, and One-hot_CNN).

By comparing the performance of different combination of three individual models and lncRNA_Mdeep, we found that the accuracy of lncRNA_Mdeep is 2.76%, 0.37%, and 1.13% higher than that of OFH_DNN + k-mer_DNN, k-mer_DNN + One-hot_CNN, and OFH_DNN + One-hot_CNN, respectively. The MCC of lncRNA_Mdeep is 0.0537, 0.0074, and 0.0222 higher than that of OFH_DNN + k-mer_DNN, k-mer_DNN + One-hot_CNN, and OFH_DNN + One-hot_CNN, respectively. These results show that lncRNA_Mdeep through fusing three models achieves better performance than that of fusing any two models.

Furthermore, we also compared lncRNA_Mdeep with a decision fusion strategy of voting. As shown in Table 1, the performance of lncRNA_Mdeep is 0.31% and 0.0059 higher than that of voting fusion strategy in terms of accuracy and MCC. All the results from Table 1 show that lncRNA_Mdeep is a superior deep learning framework and it can effectively distinguish lncRNAs from protein-coding transcripts.

Effects of different hyper-parameters

We evaluated the effects of two parameters of *k* in *k*-mer feature and *maxlen* for padding one-hot encoding. The accuracies of *k*-mer_DNN and One-hot_CNN on Human training dataset in 10CV test at different *k* and *maxlen* are shown in Figure 2. As shown in Figure 2A, we found that *k*-mer_DNN achieved

the highest accuracy when k = 6. Results in Figure 2B shows that One-hot_CNN achieves the highest accuracy when *maxlen* = 3000, Therefore, we set k = 6 when we extract the *k*-mer feature, and fix the one-hot encoding of a transcript as a 4 × *3000* matrix. All other hyper-parameters in lncRNA_Mdeep are selected by using hyperopt [35] strategy, and these parameters are listed in Additional file 2.

Comparison with other existing methods

We compared lncRNA_Mdeep with other eight existing alignment-free methods (i.e., CNCI, CPAT, PLEK, lncRNA-MEDL, CPC2, lncRNAnet, LncFinder[1] and LncFinder[2]) on human datasets and cross-species datasets. LncRNA_Mdeep is trained on Human training dataset, and since most existing methods do not provide the retraining option, we used their per-trained models.

Comparison performance on human dataset

We first compared the performance of lncRNA_Mdeep and other eight existing methods on Human testing dataset. The results are shown in Table 3, from which we can see that lncRNA_Mdeep achieves an accuracy of 93.12%, which is 6.72%, 5.14%, 15.41%, 7.65%, 15.14%, 0.94%, 6.90%, and 6.24% higher than that of CNCI, CPAT, PLEK, lncRNA-MEDL, CPC2, lncRNAnet, LncFinder[1], and LncFinder[2], respectively. MCC and S_p of lncRNA_Mdeep are 0.8653 and 88.97%, which are at least 0.0183 and 1.24% higher than that of eight methods. Although CNCI achieves 97.42% sensitivity, which is 0.15% higher than that of lncRNA_Mdeep, it shows lower performance in terms of accuracy, S_p, and MCC.

To further evaluate the “memory” effect, we trained 3 lncRNA_Mdeep models (namely model-1, model-2, and model-3) on Human training dataset, Human gene-wise training dataset, and Human non-gene-wise training dataset, respectively, and compared the performance of these 3 lncRNA_Mdeep models and other eight existing methods on Human gene-wise testing dataset. The results are shown in Additional file 3, from which we can see that lncRNA_Mdeep model-1 achieves best performance, and lncRNA_Mdeep model-2 and lncRNA_Mdeep model-3 still show the better performance than most of the existing methods. By comparing the performance of lncRNA_Mdeep model-2 with model-3, we found that lncRNA_Mdeep model-2 show the better performance than model-3. These results indicate that even if there is no overlapped originated gene between training transcripts and testing transcripts, our lncRNA_Mdeep method can still achieve a superior performance. It should be pointed out that all of other compared methods have not been retrained, and some transcripts in their training datasets may be same as the transcripts in Human gene-wise testing dataset, thus the prediction performance of these compared methods might be over-estimated.

Comparison performance on cross-species datasets

We also compared the performance of lncRNA_Mdeep and other eight existing methods by using 11 cross-species datasets as the independent testing datasets. In these tests, our lncRNA_Mdeep and other eight predictors are trained on the human dataset, and no additional training processes with other species were implemented. The results are shown in Table 4. On mouse testing dataset, lncRNA_Mdeep achieves

92.52% accuracy, which is 5.43%, 2.05%, 20.63%, 3.99%, 12.09%, 0.71%, 4.05%, and 3.53% higher than that of CNCI, CPAT, PLEK, lncRNA-MEDL, CPC2, lncRNAnet, LncFinder^{1}, and LncFinder^{2}, respectively. On other 10 cross-species testing datasets, the accuracies of lncRNA_Mdeep for arabidopsis, Bos taurus, C. elegans, chicken, chimpanzee, frog, fruit fly, gorilla, pig and zebrafish are 95.73%, 97.33%, 98.87%, 96.06%, 96.76%, 96.80%, 96.10%, 96.65%, 96.87%, and 96.76%, respectively. LncRNA_Mdeep shows the best performance on 5 out of 11 cross-species testing datasets, and lncRNA-MDFL shows the best performance on 3 cross-species testing datasets, and lncFinder^{2} shows the best performance on 3 cross-species testing datasets, and CPAT shows the best performance on 1 cross-species testing datasets. The results on the 11 testing datasets show that lncRNA_Mdeep has the superior performance for distinguishing lncRNAs from protein-coding transcripts.

Furthermore, we compared the prediction results of lncRNA_Mdeep model-1, model-2, and model-3 on 11 cross-species datasets. The results are shown in Additional file 4. By comparing the prediction results of lncRNA_Mdeep model-1, model-2, and model-3, we found that the lncRNA_Mdeep model-1 shows the best performance on 7 out of 11 cross-species datasets, and lncRNA_Mdeep model-2 shows the best performance on one cross-species dataset. By comparing the prediction results of lncRNA_Mdeep model-2 and model-3, we found that lncRNA_Mdeep model-2 achieves the better performance than model-3 on all 11 cross-species datasets. These results further show that the “memory” effect on lncRNA_Mdeep is limited, and our lncRNA_Mdeep show a better generalization performance.

Discussion

LncRNA identification is essential for understanding the lncRNA function and regulatory mechanism. In recent years, several computational methods are developed for distinguishing lncRNAs from protein-coding transcripts. Most of existing methods focused on manually extracting features and directly feeding into a classifier (e.g., support vector machine, logistic regression, and random forest) to predict lncRNAs. These predictors depend on the effectiveness of manually crafted features and fail to automatically learn intrinsic representations from raw transcript sequences. To address this issue, lncRNA_Mdeep is proposed to identify lncRNAs by multimodal deep learning. LncRNA_Mdeep can successfully integrate the manually crafted features and the raw transcript sequences. It also successfully learns high-level abstract representations based on different raw input features, and integrates learned high-level abstract representations by a multimodal deep learning model to predict lncRNAs.

Our experience results show that lncRNA_Mdeep is a superior predictor for distinguishing lncRNAs from protein-coding transcripts. We compared lncRNA_Mdeep with other different model architectures on human dataset in 10CV test, and compared lncRNA_Mdeep with existing eight state-of-the-art methods on human and 11 cross-species datasets in independent test. The results in Table 1, 3 and 4 show that lncRNA_Mdeep is a superior multimodal framework, and it achieves a better performance than other methods on human and 11 cross-species datasets. In addition, we evaluated the “memory” effect caused by the overlapped originated genes between training transcripts and testing transcripts. The results in

Table 2, Additional file 3 and 4 show that the “sequence memorization” has certain influence on the prediction results, but lncRNA_Mdeep still shows a better generalization performance.

Although lncRNA_Mdeep shows a superior performance to identify lncRNAs, there are still several issues need to be addressed in the future. First, lncRNA_Mdeep used one-hot encoding strategy to encode the raw transcript sequence and set up a parameter of *maxlen* to meet the input requirement of CNN model, but we expect a more effective encoding strategy to encode the transcript sequences with variable-length. Second, the deep learning model is like a black box, which can interpret the meaning of learned high-level abstract representations, but we expect a good way to analyze the learned high-level abstract representations.

Conclusions

In this study, we proposed a novel multimodal deep learning method (namely lncRNA_Mdeep) to distinguish lncRNAs from protein-coding transcripts. LncRNA_Mdeep first builds three individual deep model architectures to learn the hidden high-level abstract representations from three input modalities (i.e., OFH modality, *k*-mer modality, and sequence modality), and high-level representations are fused to feed into another deep model architecture for predicting lncRNAs. The experimental results show that lncRNA_Mdeep successfully integrates the manually crafted features (i.e., OFH and *k*-mer features) and the raw transcript sequences by using the multimodal framework, and it achieves higher performance than other state-of-the-art methods on human and other 11 cross-species datasets. These results indicate that lncRNA_Mdeep can contribute to the identification of novel lncRNA transcripts.

Methods

Due to technological limitations, the Methods section is only available as a download in the supplementary files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

lncRNA_Mdeep is available at: https://github.com/NWPU-903PR/lncRNA_Mdeep.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (http://www.nsfc.gov.cn/61873202, 61473232 and 91430111) awarded to SWZ. The funders did not play any role in this study.

Authors' contributions

XNF collected the dataset, performed the experiments, and wrote the initial manuscript. SWZ designed the experiments and revised the manuscript. XNF, SYZ and JJN analyzed the results. XNF developed the codes. All authors participated in the definition of the process, the discussion of relevant aspects, and approved the final manuscript.

Acknowledgements

Not applicable.

Additional files

Additional file 1: Performance of lncRNA_Mdeep and other model architectures on Human gene-wise training dataset and Human non-gene-wise training dataset in 10CV. (DOCX)

Additional file 2: The hyper-parameters in lncRNA_Mdeep. (TIF)

Additional file 3: Performance of lncRNA_Mdeep and other eight predictors on Human gene-wise testing dataset. (DOCX)

Additional file 4: Accuracy (%) of lncRNA_Mdeep model-1, model-2, model-3 and other eight predictors on 11 cross-species datasets

Additional file 5: The statistics of all datasets. (DOCX)

Abbreviations

lncRNAs: long noncoding RNAs; SVM: support vector machine; CNN: convolutional neural networks; ORF: open reading frame; RNN: recurrent neural network; DNN: deep neural network; ACC: accuracy; S_n: sensitivity; S_p: specificity; MCC: Matthew's correlation coefficient

References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al*. **Landscape of transcription in human cells**. *Nature.* 2012; **489**(7414):101-108.

2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL *et al*. **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science.* 2007; **316**(5830):1484-1488.
3. Mattick JS, Rinn JL. **Discovery and annotation of long noncoding RNAs**. *Nat Struct Mol Biol.* 2015; **22**(1):5-7.
4. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG *et al*. **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome Res.* 2012; **22**(9):1775-1789.
5. Rinn JL, Chang HY. **Genome regulation by long noncoding RNAs**. *Annu Rev Biochem.* 2012; **81**:145-166.
6. Ponting CP, Oliver PL, Reik W. **Evolution and Functions of Long Noncoding RNAs**. *Cell.* 2009; **136**(4):629-641.
7. Wapinski O, Chang HY. **Long noncoding RNAs and human disease**. *Trends Cell Biol.* 2011; **21**(6):354-361.
8. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. **CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine**. *Nucleic Acids Res.* 2007; **35**:W345-W349.
9. Lin MF, Jungreis I, Kellis M. **PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions**. *Bioinformatics.* 2011; **27**(13):I275-I282.
10. Hu L, Xu ZY, Hu BQ, Lu ZJ. **COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features**. *Nucleic Acids Res.* 2017; **45**(1):e2.
11. Achawanantakun R, Chen J, Sun YN, Zhang Y. **LncRNA-ID: Long non-coding RNA IDentification using balanced random forests**. *Bioinformatics.* 2015; **31**(24):3897-3905.
12. Sun L, Liu H, Zhang L, Meng J. **lncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine**. *Plos One.* 2015; **10**(10): e0139654.
13. 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):3825-3834.
14. Sun L, Luo HT, Bu DC, Zhao GG, Yu KT, Zhang CH, Liu YN, Chen RS, Zhao Y. **Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts**. *Nucleic Acids Res.* 2013; **41**(17):e166.
15. Wang L, Park HJ, Dasari S, Wang SQ, Kocher JP, Li W. **CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model**. *Nucleic Acids Res.* 2013; **41**(6):e74.
16. Li AM, Zhang JY, Zhou ZY. **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**.
17. Fan XN, Zhang SW. **lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning**. *Mol Biosyst.* 2015; **11**(3):892-897.

18. Tripathi R, Patel S, Kumari V, Chakraborty P, Varadwaj PK. **DeepLNC, a long non-coding RNA prediction tool using deep neural network**. Netw Model Anal Health Inform Bioinforma. 2016; **5**(1):21.
19. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei LP, Gao G. **CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features**. *Nucleic Acids Res.* 2017; **45**(W1):W12-W16.
20. Baek J, Lee B, Kwon S, Yoon S. **LncRNAnet: long non-coding RNA identification using deep learning**. *Bioinformatics.* 2018; **34**(22):3889-3897.
21. Han S, Liang Y, Ma Q, Xu Y, Zhang Y, Du W, Wang C, Li Y. **LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property**. *Brief Bioinform.* 2019; 20(6):2009-2027.
22. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang HZ, Lopez R *et al*. **The Universal Protein Resource (UniProt): an expanding universe of protein information**. *Nucleic Acids Res.* 2006; **34**:D187-D191.
23. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res.* 1997; **25**(17):3389-3402.
24. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A *et al*. **The Pfam protein families database: towards a more sustainable future**. *Nucleic Acids Res.* 2016; **44**(D1):D279-D285.
25. Finn RD, Clements J, Eddy SR. **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Res.* 2011; **39**:W29-W37.
26. Chou KC, Zhang CT. **Prediction of Protein Structural Classes**. *Crc Critical Reviews in Biochemistry.* 1995; **30**(4):275-349.
27. Chou KC. **Some remarks on protein attribute prediction and pseudo amino acid composition**. *Journal of Theoretical Biology.* 2011; **273**(1):236-247.
28. Zhang SW, Liu YF, Yu Y, hang TH, Fan XN. **MSLoc-DT: A new method for predicting the protein subcellular location of multispecies based on decision templates**. *Analytical Biochemistry.* 2014; **449**:164-171.
29. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. **iPTM-mLys: identifying multiple lysine PTM sites and their different types**. *Bioinformatics.* 2016(20):3116-3123.
30. Deng L, Wang J, Xiao Y, Wang Z, Liu H. **Accurate prediction of protein-lncRNA interactions by diffusion and HeteSim features across heterogeneous network**. *Bmc Bioinformatics.* 2018; **19(1)**:370.
31. Zhang TH, Zhang SW. **Advances in the Prediction of Protein Subcellular Locations with Machine Learning**. *Current Bioinformatics.* 2019; **14**(5):406-421.
32. Zhang SW, Fan XN. **Computational Methods for Predicting ncRNA-protein Interactions**. *Medicinal Chemistry.* 2017; 13(6):515-525.
33. Chollet F. **Keras: The python deep learning library**. 2018.

34. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M: **Tensorflow: A system for large-scale machine learning**. In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16): 2016*. 265-283.
35. Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DDJCS, Discovery. **Hyperopt: a python library for model selection and hyperparameter optimization**. 2015; **8**(1):014008.
36. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J *et al*. **GENCODE reference annotation for the human and mouse genomes**. *Nucleic Acids Res.* 2019; **47**(D1):D766-D773.
37. Pruitt KD, Tatusova T, Brown GR, Maglott DR. **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy**. *Nucleic Acids Res.* 2012; **40**(Database issue):D130-135.
38. Fickett JW. **Recognition of protein coding regions in DNA sequences**. *Nucleic Acids Res.* 1982; **10**(17):5303-5318.
39. Svozil D, Kvasnicka V, Pospichal JJC, systems il. **Introduction to multi-layer feed-forward neural networks**. 1997; **39**(1):43-62.
40. Min S, Lee B, Yoon S. **Deep learning in bioinformatics**. *Brief Bioinform.* 2017; **18**(5):851-869.
41. Hashemifar S, Neyshabur B, Khan AA, Xu J. **Predicting protein-protein interactions through sequence-based deep learning**. *Bioinformatics.* 2018; **34**(17):i802-i810.

Tables

Table 1. Performance of lncRNA_Mdeep and other model architectures on Human training dataset in 10CV test

	ACC (%)	S_n (%)	S_p (%)	MCC
OFH_DNN	95.74 ± 1.70	94.44 ± 4.89	97.04 ± 2.15	0.9171 ± 0.0307
k-mer_DNN	96.53 ± 0.41	96.40 ± 1.11	96.66 ± 0.78	0.9307 ± 0.0082
One-hot_CNN	95.82 ± 0.33	97.01 ± 0.96	94.63 ± 1.19	0.9169 ± 0.0064
OFH_DNN + k-mer_DNN	95.97 ± 2.49	96.87 ± 1.05	95.06 ± 5.71	0.9211 ± 0.0449
k-mer_DNN + One-hot_CNN	98.36 ± 0.16	98.70 ± 0.42	98.03 ± 0.50	0.9674 ± 0.0033
OFH_DNN + One-hot_CNN	97.60 ± 1.26	97.78 ± 1.58	97.43 ± 2.33	0.9526 ± 0.0248
Decision fusion	98.42 ± 1.12	**99.24** ± 0.45	97.60 ± 2.59	0.9689 ± 0.0212
lncRNA_Mdeep	**98.73** ± 0.41	98.95 ± 0.54	**98.52** ± 0.92	**0.9748** ± 0.0080

Table 2. Performance of lncRNA_Mdeep on three training datasets in 10CV

Dataset	Transcripts	ACC (%)	S_n (%)	S_p (%)	MCC
Human training dataset	46,000	**98.73** ± 0.41	**98.95** ± 0.54	**98.52** ± 0.92	**0.9748** ± 0.0080
Human non-gene-wise training dataset	24,000	97.91 ± 0.68	97.21 ± 1.40	98.61 ± 0.65	0.9584 ± 0.0133
Human gene-wise training dataset	24,000	94.95 ± 0.45	95.40 ± 1.50	94.50 ± 1.26	0.8994 ± 0.0089

Table 3. Performance of lncRNA_Mdeep and other eight methods on Human testing dataset

Methods	ACC (%)	*Sn* (%)	*Sp* (%)	MCC
CNCI	86.40	**97.42**	75.38	0.7463
CPAT	87.98	95.22	80.73	0.7676
PLEK	77.71	97.22	58.20	0.6019
lncRNA-MFDL	85.47	93.43	77.50	0.7185
CPC2	77.98	94.07	61.90	0.5911
lncRNAnet	92.18	96.63	87.73	0.8470
lncFinder[1]	86.22	95.20	77.23	0.7363
lncFinder[2]	86.88	95.98	77.77	0.7501
lncRNA_Mdeep	**93.12**	97.27	**88.97**	**0.8653**

Table 4. Accuracy (%) of lncRNA_Mdeep and other eight methods on 11 cross-species datasets

Species	CNCI	CPAT	PLEK	lncRNA-MFDL	CPC2	lncRNAnet	lncFinder[1]	lncFinder[2]	lncRNA_Mdeep
Mouse	87.09	90.47	71.89	88.53	80.43	91.81	88.47	88.99	**92.52**
Arabidopsis	79.86	91.39	66.93	**97.30**	93.36	94.60	92.45	93.77	95.73
Bos taurus	92.88	97.13	89.32	95.51	96.10	96.30	97.00	97.03	**97.33**
C. elegans	77.72	91.48	45.37	97.97	94.75	97.95	87.46	88.55	**98.87**
Chicken	91.52	**97.04**	83.95	96.87	95.22	95.56	96.82	96.64	96.06
Chimpanzee	89.84	96.18	88.99	94.26	95.48	94.78	96.05	96.21	**96.76**
Frog	90.60	96.40	80.90	96.14	96.34	95.53	96.92	**97.26**	96.80
Fruit fly	92.90	96.02	74.43	**96.49**	94.28	95.21	95.33	95.50	96.10
Gorilla	89.37	94.99	86.75	95.12	94.12	94.31	94.72	94.87	**95.65**
Pig	91.73	96.91	87.34	**96.98**	95.86	95.56	96.88	96.82	96.87
Zebrafish	93.59	97.50	85.07	92.17	96.83	95.77	97.54	**97.78**	96.76

Figures

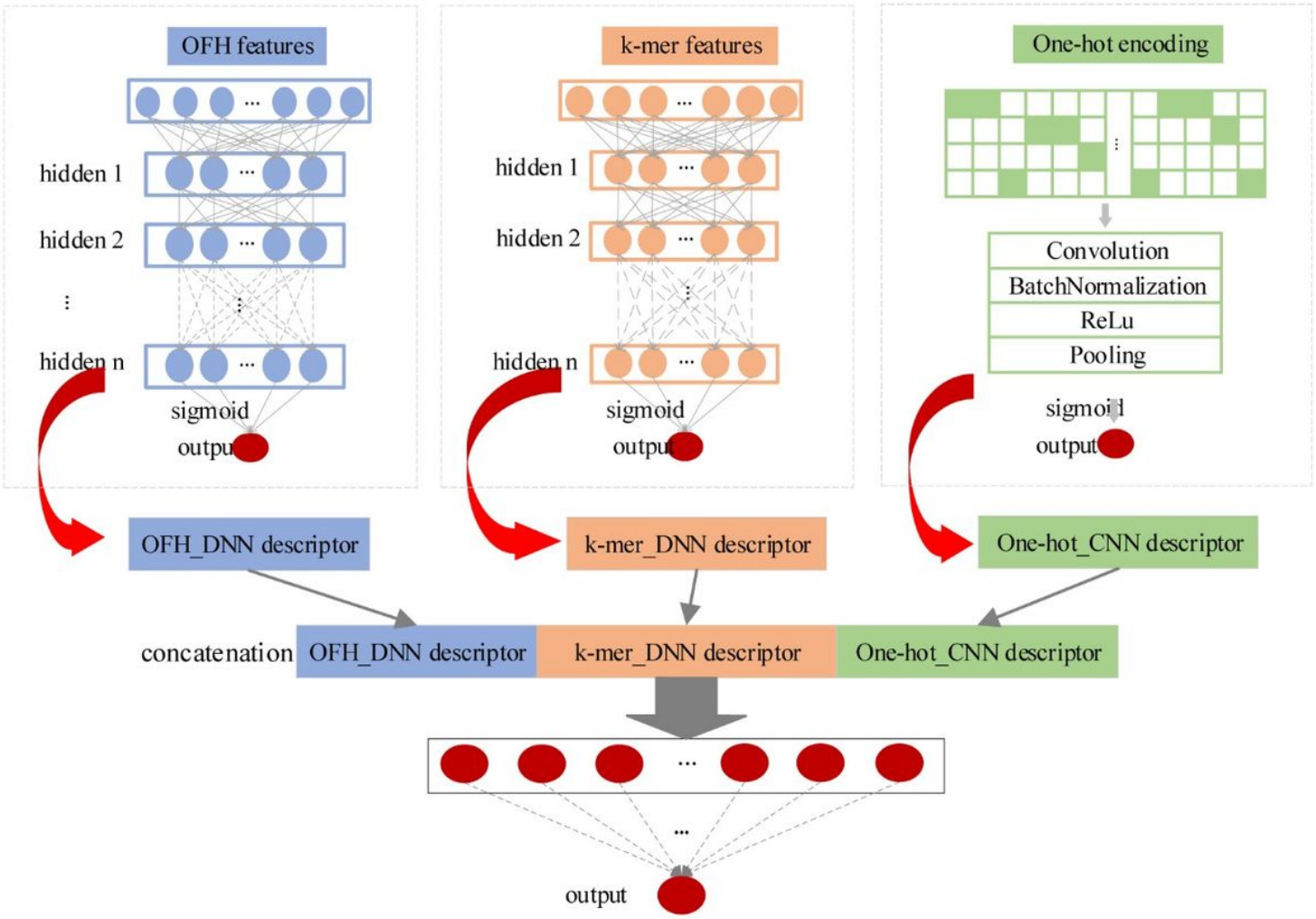

Figure 1

Overview of lncRNA_Mdeep

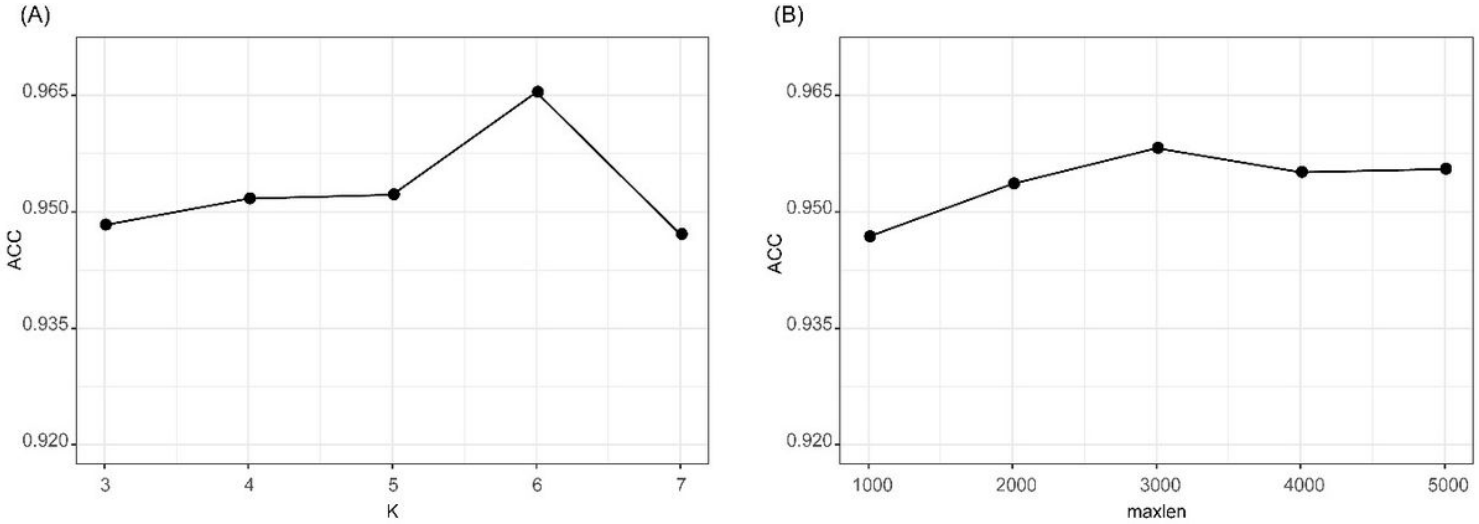

Figure 2

Results of k-mer_DNN and One-hot_CNN with different parameters. (A) Accuracy of k-mer_DNN with different k value. (B) Accuracy of One-hot_CNN with different maxlen value.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RevisedAdditionalfile4.docx
- RevisedAdditionalfile2.tif
- Methods.docx
- RevisedAdditionalfile5.docx
- RevisedAdditionalfile3.docx
- RevisedAdditionalfile1.docx