piRNN: deep learning algorithm for piRNA prediction

Kai Wang[1], Joshua Hoeksema[2] and Chun Liang[1]

[1] Department of Biology, Miami University, Oxford, OH, USA
[2] Department of Computer Science & Software Engineering, Miami University, Oxford, OH, USA

ABSTRACT

Piwi-interacting RNAs (piRNAs) are the largest class of small non-coding RNAs discovered in germ cells. Identifying piRNAs from small RNA data is a challenging task due to the lack of conserved sequences and structural features of piRNAs. Many programs have been developed to identify piRNA from small RNA data. However, these programs have limitations. They either rely on extracting complicated features, or only demonstrate strong performance on transposon related piRNAs. Here we proposed a new program called piRNN for piRNA identification. For our software, we applied a convolutional neural network classifier that was trained on the datasets from four different species (*Caenorhabditis elegans*, *Drosophila melanogaster*, rat and human). A matrix of *k-mer* frequency values was used to represent each sequence. piRNN has great usability and shows better performance in comparison with other programs. It is freely available at https://github.com/bioinfolabmu/piRNN.

Subjects Bioinformatics, Computational Biology, Data Mining and Machine Learning
Keywords piRNA, Deep learning, Convolution neural network

Submitted 25 April 2018
Accepted 19 July 2018
Published 3 August 2018

Corresponding author
Chun Liang, liangc@miamioh.edu

Academic editor
Lourdes Peña-Castillo

Additional Information and Declarations can be found on page 8

DOI **10.7717/peerj.5429**

INTRODUCTION

Piwi-interacting RNAs (piRNAs) are the largest class of small non-coding RNAs that are enriched in animal germline cells (*Lau et al., 2006*). piRNAs were thought to be transposon silencers in germline cells to maintain genome integrity through epigenetic and post-transcriptional regulations (*Aravin, Hannon & Brennecke, 2007*). Recent studies reveal that piRNAs are found in somatic cells and are also related to certain mRNAs degradation (*Malone et al., 2009*; *Rouget et al., 2010*). This suggests that piRNAs may have more important functions in germline and somatic cells that are waiting for further exploration. The first step in studying the function of piRNAs is to identify them from other small non-coding RNAs like miRNA, siRNA, etc. Experimental approaches like extracting piRNAs or obtaining transposon-related small RNAs from germline cells have been applied to identify piRNAs (*Aravin et al., 2007*; *Lau et al., 2006*). Since piRNAs are not only expressed in germline cells and their functions are not limited to silencing transposons, those experimental approaches have limitations (e.g., incapability of capturing piRNAs with low expression levels and exploring somatic piRNAs). Therefore, computational approaches were proposed to predict piRNAs from small RNA sequence data (*Betel et al., 2007*).

How to cite this article Wang et al. (2018), piRNN: deep learning algorithm for piRNA prediction. **PeerJ 6:e5429; DOI 10.7717/peerj.5429**

The first bioinformatics tool for piRNA prediction was created by applying Fisher discriminant algorithm to *k-mer* sequence features using small RNA data (*Zhang, Wang & Kang, 2011*). This algorithm can achieve a precision of >90% and a sensitivity of >60% in five species data (nematode, fruit fly, human, rat, and mouse). Piano (*Wang et al., 2014*) and piRNAPredictor (*Li et al., 2016*) were developed to predict transposon-related piRNAs by using support vector machine and weighted ensemble method, and they achieved a very high accuracy of over 95%. Another transposon-related piRNA prediction program that developed using ensemble learning method also achieved great performance (*Luo et al., 2016*). A previous study (*Beyret, Liu & Lin, 2012*) suggested that only a portion of piRNAs are related to transposons, so the aforementioned three programs are only suitable for a subset of total piRNAs whereas a large portion of piRNAs would be ignored by these programs. In addition, other programs were also developed to predict total piRNAs from small RNA data. piRPred (*Brayet et al., 2014*) used a support vector machine and kernel combination strategy to achieve over 86% accuracy on human data and over 89% accuracy on fruit fly data. In piRPred, sequence information (i.e., *k-mer* and 5′ uridine) and genome information (i.e., distance to pericentromeric and subtelomeric regions and piRNA cluster location) were extracted and combined together as sequence features for SVM training and testing. 2L-piRNA (*Liu, Yang & Chou, 2017*) is a two-layer ensemble classifier for piRNA prediction and functional annotation, with the accuracy of 86.1%. Since the accuracy/specificity of piRPred and 2L-piRNA is around 85%, there is still room for improvement in piRNA prediction. Pibomd (*Liu, Ding & Gong, 2014*) is also a piRNA prediction program established on a support vector machine by using sequence motif features. The sensitivity and specificity of Pibomd are around 90%. Since Pibomd is trained only on mouse data, it cannot be used for other species. IpiRId (*Boucheham et al., 2017*) combined sequence content features and epigenomic features in piRNA prediction and achieved around 90% accuracy on human, mouse, and fruit fly data. However, its feature extraction is complicated and the epigenomic data in addition to small RNA data is also needed, making IpiRId difficult to apply on non-model organisms. V-ELMpiRNAPred (*Pian et al., 2017*) was developed for human piRNA prediction with a high specificity(95%) and sensitivity (94%). Unfortunately, it is difficult to apply V-ELMpiRNAPred to other model organisms.

Recently, deep learning is a very popular and widely used technique for many kinds of classification tasks (*LeCun, Bengio & Hinton, 2015*). Deep learning algorithms have been utilized for several bioinformatics applications. For example, DeepBind was developed for predicting DNA- and RNA-binding proteins (*Alipanahi et al., 2015*) and DeepChrome for predicting gene expression levels from histone modification data (*Singh et al., 2016*). In this paper, we propose a new program called piRNN for piRNA prediction that was developed by using a deep learning model based on convolution neural network (CNN) framework. Our program achieved over 90% accuracy on *Caenorhabditis elegans*, *Drosophila melanogaster*, rat and human data.

In order to develop a powerful piRNA prediction program for biologists to find piRNA accurately from small RNA data, there are five critical questions need to be addressed properly as follows (*Chou, 2011*): How to select valid datasets to train and test the predictor; how to use effective mathematical expression to represent biological sequences; how to develop a powerful algorithm to operate the prediction; how to perform cross-validation tests properly to evaluate the predictor; and how to make the predictor user-friendly and can be easily accessed by the public? In piRNN, the above five questions are adequately addressed and described in details in the following parts.

MATERIALS AND METHODS

Data sets

In our program, *C. elegans*, *D. melanogaster*, rat and human piRNAs were used for training models. *C. elegans*, rat and human data were downloaded from piRBase (*Zhang et al., 2014*). *D. melanogaster* data were downloaded from piRNABank (*Lakshmi & Agrawal, 2008*). In total, 28,219 *C. elegans* piRNA sequences, 22,336 *D. melanogaster* piRNA sequences, 63,182 rat piRNA sequences, and 32,826 human piRNA sequences were used as the positive datasets. To develop a powerful DNA/RNA sequence classifier, one of the most critical and challenging problems is how to generate a valid negative dataset that can be used effectively in model training and testing. Previous researches such as PseKNC (*Chen et al., 2014*) and Pse-in-One (*Liu et al., 2015*) provided comprehensive methods that can be used to generate pseudo DNA/RNA sequences (i.e., negative dataset) for computational biology research. Based on the previous study (*Brayet et al., 2014*), the negative datasets of each species consist of three parts: (1) corresponding fake-piRNA sequences generated using the first order Markov model where the probability distribution came from the positive datasets, (2) corresponding mature microRNA sequences downloaded from miRBase (*Kozomara & Griffiths-Jones, 2014*), and (3) tRNA fragments of certain lengths randomly cut from tRNA sequences downloaded from Genomic tRNA database (*Chan & Lowe, 2009*). Accordingly, both positive and negative datasets have the same sequence numbers, respectively, for each species. Each species' dataset, both positive and negative data, was divided into two parts: training-testing dataset (90%) and validation dataset (10%). All models were trained using 10-fold cross validation with the training-testing data set. The validation dataset was used for comparison with other programs. More details can be found in the Supplemental File.

Sequence feature extraction

Previous studies demonstrate that sequence-derived features can be efficiently used for developing different types of DNA/RNA sequence classification programs (*Zhang et al., 2017a*, *2017b*). In our program, the sequence feature consists of two parts. The first part is *k-mer* ($k = 1, 2, 3, 4, 5$) motif frequencies. In total, 1,364 (sum of 4^1, 4^2, 4^3, 4^4, and 4^5) values were extracted from each sequence. For the second part, we counted the *k-mer* motifs around the first and 10th base because piRNA has a high probability starting with a "T/U" and containing an "A" at 10th position (*Hirakata & Siomi, 2016*). Specifically, if the sequence starts with a "T/U" and/or has an "A" in the 10th position, then we updated the second part feature vectors into the first vectors as shown in Fig. 1. For training, the total 1,364 vectors were transformed into a 4 × 341 matrix (Fig. 1).

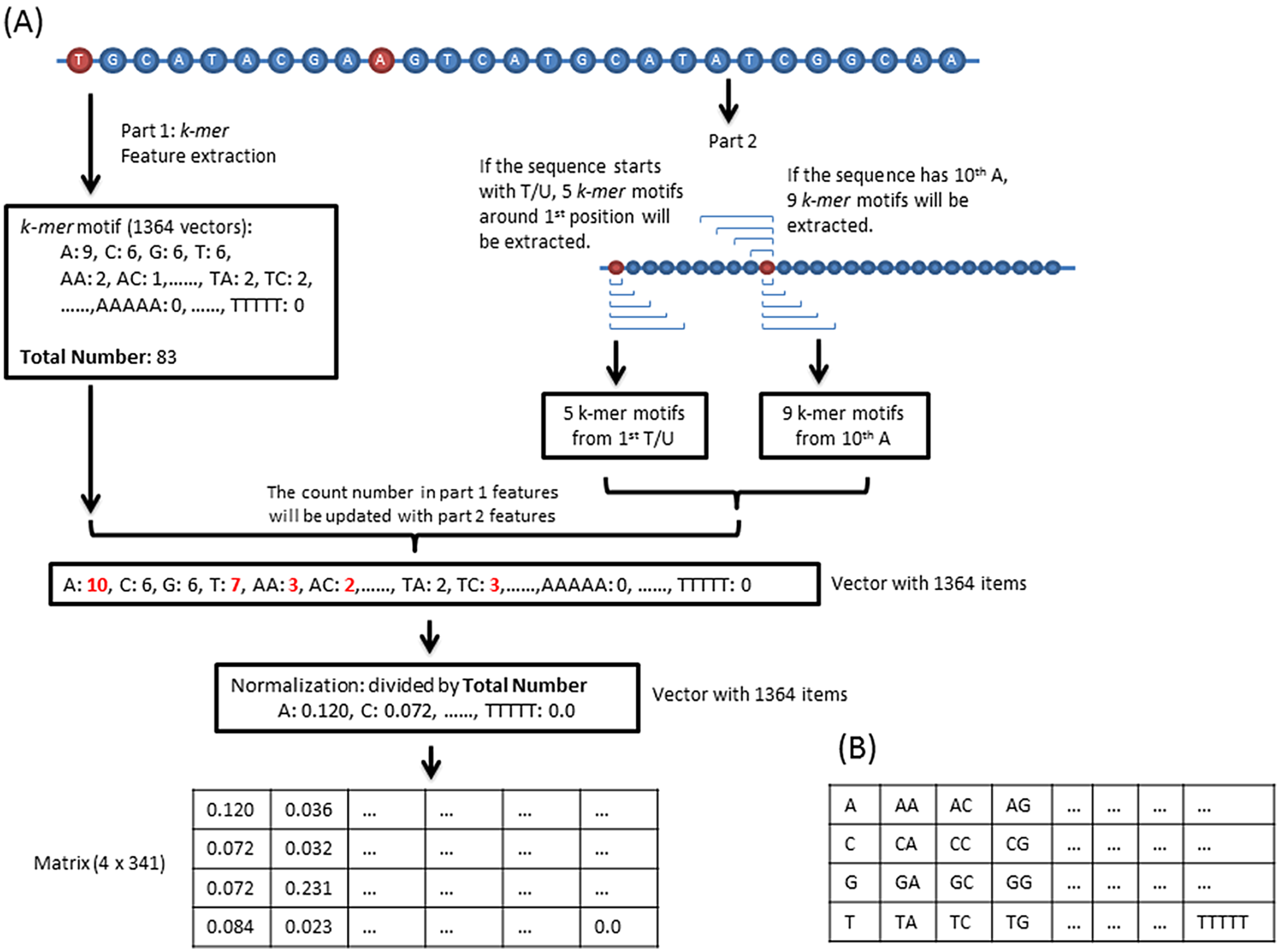
(A)
T G C A T A C G A A G T C A T G C A T A T C G G C A A
Part 1: k-mer
Feature extraction
k-mer motif (1364 vectors):
A: 9, C: 6, G: 6, T: 6,
AA: 2, AC: 1,......, TA: 2, TC: 2,
......,AAAAA: 0, ..., TTTTT: 0
Total Number: 83
Part 2
If the sequence starts with T/U, 5 k-mer motifs around 1st position will be extracted.
If the sequence has 10th A, 9 k-mer motifs will be extracted.
5 k-mer motifs from 1st T/U
9 k-mer motifs from 10th A
The count number in part 1 features will be updated with part 2 features
A: 10, C: 6, G: 6, T: 7, AA: 3, AC: 2,......, TA: 2, TC: 3,......,AAAAA: 0, ..., TTTTT: 0
Vector with 1364 items
Normalization: divided by Total Number
A: 0.120, C: 0.072, ..., TTTTT: 0.0
Vector with 1364 items
Matrix (4 x 341)
0.120 0.036 ... ...
0.072 0.032 ... ...
0.072 0.231 ... ...
0.084 0.023 ... ... 0.0
(B)
A AA AC AG ... ...
C CA CC CG ... ...
G GA GC GG ... ...
T TA TC TG ... ... TTTTT

Figure 1 **Feature extraction.** (A) Sequence features consist of two parts. The first part is the *k-mer* motifs of the whole sequence. The second part *k-mer* motifs are around the first T/U and/or 10th A. If a sequence dose not start with a T/U or do not have a 10th A, the corresponding second part will not be calculated. (B) The 1,364 vectors are transformed into this matrix. Full-size DOI: 10.7717/peerj.5429/fig-1

CNN architecture

A CNN is a feed-forward artificial neural network that has been widely used in image recognition. In this paper, we have implemented a CNN for piRNA identification by using TensorFlow (*Abadi et al., 2016*) and Keras (*Chollet, 2015*). The whole program was implemented in Python 3.5. The architecture of piRNN is showed in Fig. 2. The input of each sequence is a 4 × 341 matrix. Each element of this matrix represents *k-mer* motifs frequencies as shown in Fig. 1. Two one-hot vectors ([0, 1], [1, 0]) were used to label the piRNA and non-piRNA sequences. The first part of our network is two convolution layers. A total of 32 2 × 2 filters were used in these two convolution layers to scan the input matrixes. The second part is a max pooling layer with 2 × 2 pooling size. A dropout layer with 0.25 dropout ratio is connected to the max pooling layer. The third part

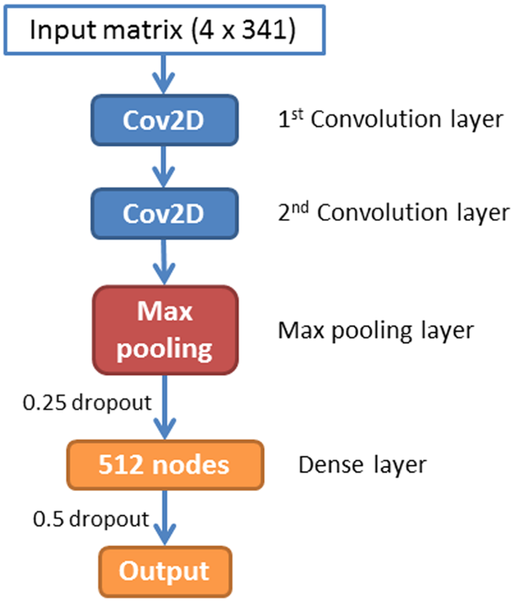

Figure 2 Architecture of the convolution neural network. Full-size DOI: 10.7717/peerj.5429/fig-2

is a dense layer with 512 nodes. We applied a 50% dropout after the dense layer to prevent over fitting. The last layer is the outputs layer with two nodes that correspond to the piRNAs and non-piRNAs.

Model training

A total of 10-fold cross validation was used for model training in four species individually. A small batch size of 32 was applied in all training processes. During the 10-fold training for all four species, the training process will be stopped if the value of cost function did not decrease after 10 epochs. Since the input of our program is a small matrix (4 × 341) and it is easy to get overfitting in CNN training, we used a simple CNN architecture that only contains two convolution layers, one max pooling layer, and one dense layer, with two dropout layers (see Fig. 2). All parameters used in training were set in terms of the default values from Keras document (https://keras.io/). The four species models were trained by using the same CNN architecture because we are aiming to establish a piRNA prediction tool that can be easily used by biologists. It is worth mentioning that piRNN allows re-training to get better models if more data is available.

RESULTS

Classification performance

After training, four species–specific models are tested using their relevant test datasets to evaluate model performance. Figure 3 shows the performance of piRNN. Accuracy (ACC), precision (*Pre*), sensitivity (*Sn*), specificity (*Sp*), and Matthews correlation coefficient (MCC) were used for performance evaluation. These measurements were calculated from the testing datasets by 10-fold cross validation. Since piRNA is small non-coding RNA only with ~30 nt in length, a large number of *k* in feature extraction

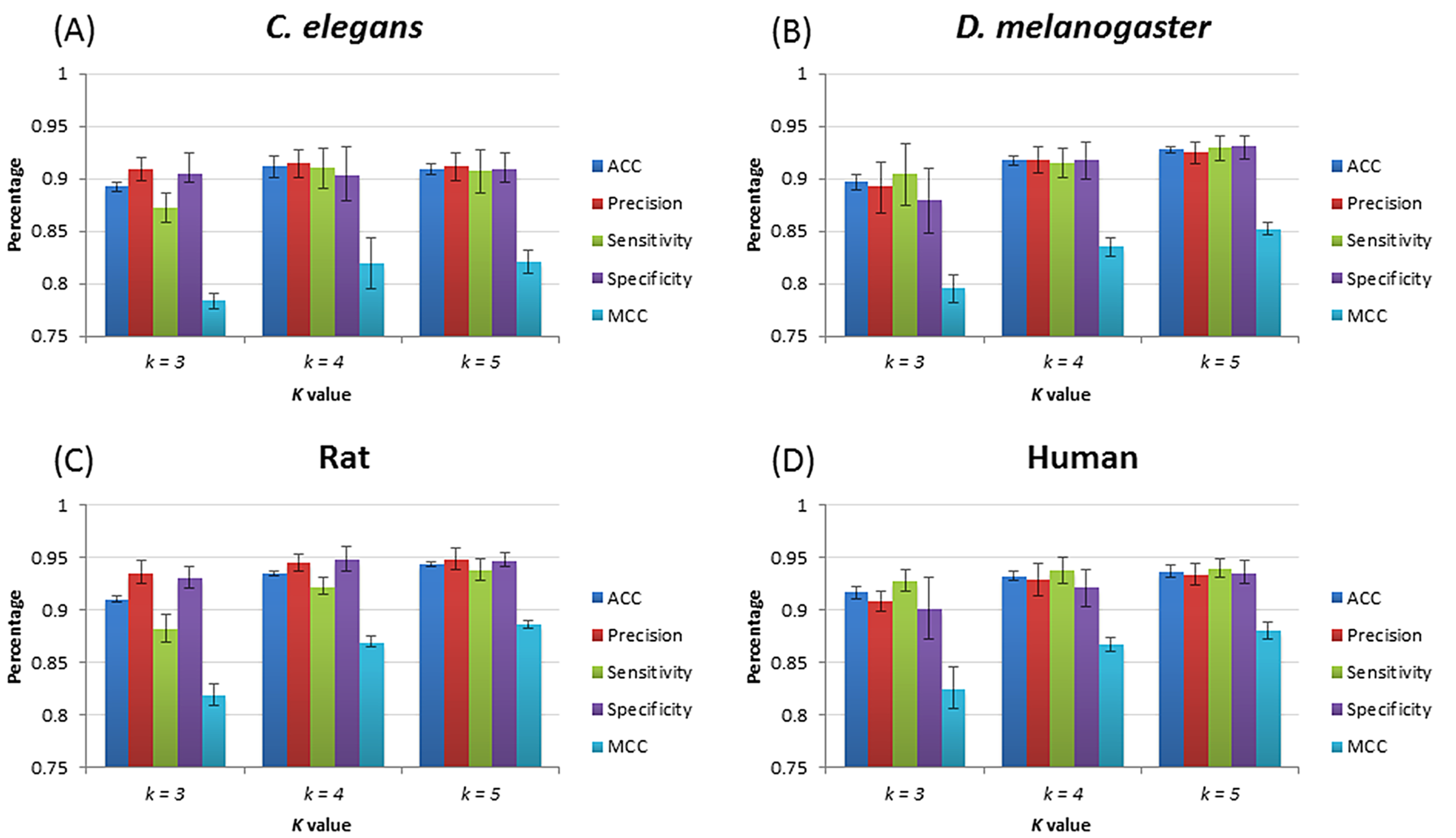

Figure 3 **piRNN performance with different *k* value for different species.** (A), (B), (C), and (D) show the results of *C. elegans*, *D. melanogaster*, rat, and human, respectively.
Full-size DOI: 10.7717/peerj.5429/fig-3

would produce many useless features (value equals 0). Based on the testing results and taking into account the calculation cost and operation time, we selected *k* = 5 lastly for our program.

Comparison with other methods

Previous piRNA prediction/identification programs can be classified into two categories or groups. The first one is those that only focused on predicting transposon-related piRNAs, such as piRNAPredictor (*Li et al., 2016*), Piano (*Wang et al., 2014*), etc. These programs usually can achieve very high accuracy and precision. Since the transposon-related piRNAs are only a small portion of total piRNAs, especially in mammals, it is easy to train a model with good performance on small datasets. In this context, non-transposon-related piRNAs (over half of total piRNAs) would be excluded from the prediction results (*Beyret, Liu & Lin, 2012*). Since piRNAs are also showed to silence certain mRNAs (*Rouget et al., 2010*), these programs have obvious limitation in piRNA prediction. The second group of programs can be used to predict total piRNAs from small RNA data. Example include piRNApredictor (*Zhang, Wang & Kang, 2011*), piRPred (*Brayet et al., 2014*), and IpiRId (*Boucheham et al., 2017*). Combined with genomic features like piRNA genome locations and epigenetic data, piRPred and

Table 1 **Comparison between Piano, piRNApredictor, 2L-piRNA, and piRNN.** The results of piRNN are highlighted in bold.

Program	Species	ACC	*Pre*	*Sn*	*Sp*	MCC
Piano	Fruit fly	0.68 ± 0.021	0.63 ± 0.018	0.87 ± 0.021	0.50 ± 0.034	0.40 ± 0.043
	Human	0.62 ± 0.013	0.58 ± 0.015	0.92 ± 0.008	0.32 ± 0.016	0.30 ± 0.02
piRNApredictor	Fruit fly	0.53 ± 0.013	0.66 ± 0.066	0.14 ± 0.025	0.93 ± 0.012	0.11 ± 0.043
	Human	0.72 ± 0.019	0.84 ± 0.019	0.55 ± 0.04	0.89 ± 0.014	0.47 ± 0.032
2L-piRNA	Fruit fly	0.52 ± 0.027	0.65 ± 0.039	0.39 ± 0.035	0.71 ± 0.30	0.10 ± 0.051
	Human	0.67 ± 0.028	0.68 ± 0.031	0.79 ± 0.025	0.51 ± 0.042	0.31 ± 0.055
piRNN	Fruit fly	**0.95 ± 0.003**	**0.93 ± 0.006**	**0.97 ± 0.004**	**0.97 ± 0.004**	**0.90 ± 0.006**
	Human	**0.95 ± 0.004**	**0.94 ± 0.008**	**0.97 ± 0.012**	**0.97 ± 0.012**	**0.91 ± 0.008**

Note:
ACC, Accuracy; *Pre,* precision; *Sn,* sensitivity; *Sp,* specificity; MCC, Matthews correlation coefficient.

IpiRId achieved better performance than the piRNApredictor in piRNA identification. Since they need genomic features for prediction, it is hard to apply these programs to non-model organisms that lack enough genomics resources.

For performance comparison, we compared our program with Piano (*Wang et al., 2014*), 2L-piRNA (*Liu, Yang & Chou, 2017*), and piRNApredictor (*Zhang, Wang & Kang, 2011*). Piano is the first piRNA prediction program that only focused on transposon-related piRNAs. piRNApredictor is the first software that can predict piRNAs using small RNA-Seq data. piRPred (*Brayet et al., 2014*), Pibomd (*Liu, Ding & Gong, 2014*), V-ELMpiRNAPred (*Pian et al., 2017*), and IpiRId (*Boucheham et al., 2017*) are web-based tools that are unable to execute locally. piRNAPredictor (*Li et al., 2016*) does not provide any standalone downloadable programs. Therefore, we only selected Piano, 2L-piRNA, and piRNApredictor for the performance comparison. A pervious study shows that *D. melanogaster* has different length distribution from mammal piRNAs (*Wang et al., 2014*), so we only used human data that is representative of mammal piRNAs and *D. melanogaster* piRNAs for comparison. Our comparison results are listed in Table 1. piRNN shows the best performance for all five measurements for both species, with accuracy, precision, sensitivity, specificity, and Matthews correlation coefficient of over 90% (see Table 1). Compared to the results provided in the paper of IpiRId (accuracy for human and fruit fly is 90.09 ± 0.25 and 92.59 ± 1.87, respectively), our program piRNN also demonstrated better performance results than IpiRId on both fruit fly and human data. Considering the fact that our program is the one with the best performance and only utilizes *k-mer* features without other data (e.g., genome location or epigenetic data) in piRNA prediction, piRNN is more user-friendly for biologists to apply on new data. Clearly, piRNN is a useful tool for piRNA prediction in non-model organisms where genomics resources are limited.

CONCLUSIONS

In summary, we developed a deep learning program for identifying piRNAs based on CNN. The major advantages of this program can be concluded as: (1) this is the first deep learning based piRNA identification program that demonstrates the best performance in comparison to three previous programs, (2) this program adopts a genome-independent approach that does not need genome and/or epigenomic data for identifying piRNAs, (3) our program is an easy-to-use program with an option to choose one of the four well-trained models for four different species (*C. elegans*, *D. melanogaster*, rat and human), and (4) we also provide the training protocol and procedure for users if they want to retrain existing models or train new models for new species and conduct predictions. Our program is freely available at https://github.com/bioinfolabmu/piRNN. User can download all source codes and testing datasets from the aforementioned website. As pointed out in recent publications (*Chou & Shen, 2009*; *Lin et al., 2014*; *Song et al., 2018*), user-friendly web-service represents the future direction for computational tools. In future, we would like to establish a web-service for piRNN so that it is within the reach of general biologists. All updates of our program will be released in our github website. We are committed to improve piRNN for better piRNA prediction for more species in the future.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This project is funded by the Committee on Faculty Research (CFR) Program, the Office for the Advancement of Research & Scholarship (OARS) and Biology Department, Miami University, Oxford, OH, USA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Committee on Faculty Research (CFR) Program.
Office for the Advancement of Research & Scholarship (OARS) and Biology Department, Miami University, Oxford, OH, USA.

Competing Interests

Chun Liang is an Academic Editor for PeerJ. The authors declare that they have no competing interests.

Author Contributions

- Kai Wang conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper.
- Joshua Hoeksema analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper.
- Chun Liang conceived and designed the experiments, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

GitHub: https://github.com/bioinfolabmu/piRNN.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.5429#supplemental-information.

REFERENCES

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M. 2016. TensorFlow: a system for large-scale machine learning. *OSDI* **16**:265–283.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33(8)**:831–838 DOI 10.1038/nbt.3300.

Aravin AA, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318(5851)**:761–764 DOI 10.1126/science.1146484.

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316(5825)**:744–747 DOI 10.1126/science.1142612.

Betel D, Sheridan R, Marks DS, Sander C. 2007. Computational analysis of mouse piRNA sequence and biogenesis. *PLOS Computational Biology* **3(11)**:e222 DOI 10.1371/journal.pcbi.0030222.

Beyret E, Liu N, Lin HF. 2012. piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. *Cell Research* **22(10)**:1429–1439 DOI 10.1038/cr.2012.120.

Boucheham A, Sommard V, Zehraoui F, Boualem A, Batouche M, Bendahmane A, Israeli D, Tahi F. 2017. IpiRId: integrative approach for piRNA prediction using genomic and epigenomic data. *PLOS ONE* **12(6)**:e0179787.

Brayet J, Zehraoui F, Jeanson-Leh L, Israeli D, Tahi F. 2014. Towards a piRNA prediction using multiple kernel fusion and support vector machine. *Bioinformatics* **30(17)**:i364–i370 DOI 10.1093/bioinformatics/btu441.

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* **37(Database)**:D93–D97 DOI 10.1093/nar/gkn787.

Chen W, Lei TY, Jin DC, Lin H, Chou KC. 2014. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Analytical Biochemistry* **456**:53–60 DOI 10.1016/j.ab.2014.04.001.

Chollet F. 2015. Keras. *Available at https://keras.io* (accessed 31 July 2018).

Chou KC. 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology* **273(1)**:236–247 DOI 10.1016/j.jtbi.2010.12.024.

Chou K-C, Shen H-B. 2009. Recent advances in developing web-servers for predicting protein attributes. *Natural Science* **1(2)**:63–92 DOI 10.4236/ns.2009.12011.

Hirakata S, Siomi MC. 2016. piRNA biogenesis in the germline: from transcription of piRNA genomic sources to piRNA maturation. *Biochimica et Biophysica Acta—Gene Regulatory Mechanisms* **1859(1)**:82–92 DOI 10.1016/j.bbagrm.2015.09.002.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* **42(D1)**:D68–D73 DOI 10.1093/nar/gkt1181.

Lakshmi SS, Agrawal S. 2008. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Research* **36(suppl_1)**:D173–D177 DOI 10.1093/nar/gkm696.

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006. Characterization of the piRNA complex from rat testes. *Science* **313(5785)**:363–367 DOI 10.1126/science.1130164.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521(7553)**:436–444 DOI 10.1038/nature14539.

Li DF, Luo LQ, Zhang W, Liu F, Luo F. 2016. A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics* **17(1)**:329 DOI 10.1186/s12859-016-1206-3.

Lin H, Deng EZ, Ding H, Chen W, Chou KC. 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Research* **42(21)**:12961–12972 DOI 10.1093/nar/gku1019.

Liu XQ, Ding J, Gong FZ. 2014. piRNA identification based on motif discovery. *Molecular BioSystems* **10(12)**:3075–3080 DOI 10.1039/c4mb00447g.

Liu B, Liu FL, Wang XL, Chen JJ, Fang LY, Chou KC. 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research* **43(W1)**:W65–W71 DOI 10.1093/nar/gkv458.

Liu B, Yang F, Chou KC. 2017. 2L-piRNA: a two-layer ensemble classifier for identifying Piwi-interacting RNAs and their function. *Molecular Therapy—Nucleic Acids* **7**:267–277 DOI 10.1016/j.omtn.2017.04.008.

Luo L, Li D, Zhang W, Tu S, Zhu X, Tian G. 2016. Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. *PLOS ONE* **11(4)**:e0153268.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* **137(3)**:522–535 DOI 10.1016/j.cell.2009.03.040.

Pian C, Chen YY, Zhang J, Chen Z, Zhang GL, Li Q, Yang T, Zhang LY. 2017. V-ELMpiRNAPred: identification of human piRNAs by the voting-based extreme learning machine (V-ELM) with a new hybrid feature. *Journal of Bioinformatics and Computational Biology* **15(1)**:1650046 DOI 10.1142/s0219720016500463.

Rouget C, Papin C, Boureux A, Meunier AC, Franco B, Robine N, Lai EC, Pelisson A, Simonelig M. 2010. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. *Nature* **467(7319)**:1128–1132 DOI 10.1038/nature09465.

Singh R, Lanchantin J, Robins G, Qi YJ. 2016. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32(17)**:i639–i648 DOI 10.1093/bioinformatics/btw427.

Song J, Wang Y, Li F, Akutsu T, Rawlings ND, Webb GI, Chou KC. 2018. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Briefings in Bioinformatics* **2018**:bby028 DOI 10.1093/bib/bby028.

Wang K, Liang C, Liu JD, Xiao HM, Huang SQ, Xu JH, Li F. 2014. Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics* **15(1)**:419 DOI 10.1186/s12859-014-0419-6.

Zhang W, Shi J, Tang G, Wu W, Yue X, Li D. 2017a. Predicting small RNAs in bacteria via sequence learning ensemble method. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Piscataway: IEEE, 643–647.

Zhang P, Si XH, Skogerbo G, Wang JJ, Cui DY, Li YX, Sun XB, Liu L, Sun BF, Chen RS, He SM, Huang DW. 2014. piRBase: a web resource assisting piRNA functional study. *Database* **2014**:bau110 DOI 10.1093/database/bau110.

Zhang W, Zhu X, Fu Y, Tsuji J, Weng Z. 2017b. Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. *BMC Bioinformatics* **18(Suppl 13)**:464 DOI 10.1186/s12859-017-1875-6.

Zhang Y, Wang X, Kang L. 2011. A *k*-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics* **27(6)**:771–776 DOI 10.1093/bioinformatics/btr016.