A MiRNA Target Prediction Model based on Distributed Representation Learning and Deep Learning

Yuzhuo Sun
Southwest Forestry University
Fei Xiong
Southwest Forestry University
Yongke Sun
Southwest Forestry University
Youjie Zhao
Southwest Forestry University
Yong Cao (cn_caoyong@126.com)
Southwest Forestry University

Research Article

Keywords: miRNAs target prediction, Distributed representation learning, Deep learning, BiLSTM, Word2vec

Posted Date: September 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-888752/v1

Version of Record: A version of this preprint was published at Computational and Mathematical Methods in Medicine on July 25th, 2022. See the published version at https://doi.org/10.1155/2022/4490154.

RESEARCH

A miRNA target prediction model based on distributed representation learning and deep learning

Yuzhuo Sun[1], Fei Xiong[1], Yongke Sun[2], Youjie Zhao[1] and Yong Cao[1*]

*Correspondence: cn_caoyong@126.com
[1]College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, China
Full list of author information is available at the end of the article

Abstract

Background: MicroRNAs (miRNAs) are a kind of non-coding RNA, which plays an essential role in gene regulation by binding to messenger RNAs(mRNAs). Accurate and rapid identification of miRNA target genes is helpful to reveal the mechanism of transcriptome regulation, which is of great significance for the study of cancer and other diseases. Many bioinformatics methods have been proposed to solve this problem, but the previous research did not further study the encoding of the base sequence.

Results: In this study, we developed a novel method combining word embedding and deep learning for human miRNA targets at the site level prediction, which is inspired by the similarity between natural language and biological sequences. First, the wor2vec model was used to mine the distribution representation of miRNAs and mRNAs. Then, the data features are fully extracted automatically from temporal and spatial via the stacked Bidirectional Long short-term memory(BiLSTM) network. We compare the effects of different embedding methods on model accuracy in different deep learning models, and the results prove that using word2vec can improve the accuracy of deep learning models. In addition, we performed visual analysis on the distributed represented sequences and found hidden similarity relationships between bases. Finally, compared with different advanced methods and data sets, the results show that our proposed method has gotten better performance in multiple evaluation aspects.

Conclusions: We present a novel method for predicting miRNA target sites consisting of word2vec and the BiLSTM model and demonstrate that this method can realize automatic feature extraction and has higher accuracy. Furthermore, we process miRNA and mRNA as two languages for the first time and explore their biological significance through visual analysis.

Keywords: miRNAs target prediction; Distributed representation learning; Deep learning; BiLSTM; Word2vec

Content

Text and results for this section, as per the individual journal's instructions for authors.

Background

miRNAs[1] are small single-stranded RNA molecules with a length of 22 nucleotides (nts), which are widely found in eukaryotes. Mature miRNAs combine with proteins to form RNA-induced silencing complexes (RISC)[2] that cause mRNA hydrolysis or inhibit translation by binding to the target sites of mRNA. miRNAs regulate more than 60% of protein-coding genes in humans and other mammals and play crucial roles in many biological processes, including cell development, differentiation, proliferation, and apoptosis[3]. Past evidence has shown that miRNAs are also closely associated with diseases such as cancer and metabolic abnormalities[4, 5]. However, up to now, the functions of a large number of miRNAs are still unclear. Therefore, finding the target sites of miRNA is of great significance for understanding its function and regulatory mechanism.

For miRNA target gene research, there are currently three types of methods that can effectively find the target sites of miRNA, but there are still some problems to be improved. The method based on the biological experiment[6] can find target genes accurately, but the experiment cycle is long and the cost is expensive. Although the method based on database[7] search and matching can get the result quickly, it can not determine the information not included in the database, and the low accuracy. A result of the above method problems prompted the development of machine learning algorithm tools. With the continuous development of artificial intelligence technology, most of the recent methods are based on deep learning. DeepTarget[8] is an end-to-end model at the two levels of processing site and gene. The feature extraction of miRNA and mRNA is carried out by autoencoder respectively, and then uses gate recurrent unit(GRU) to learn the sequence-to-sequence interactions between miRNA and their targets. Deepmirtar[9] uses 750 manually extracted features in 7 categories, using a stack denoising autoencoder as a model, and achieves 93% accuracy at the site level. Xueming[10] uses a multi-layer convolutional neural network(CNN) stack structure, processing site, and gene rank prediction, and the model can use full-length mRNAs as input. However, even though these deep learning methods have achieved good results, there are still some problems that need to be improved. In the past, the traditional methods used the simulation method to generate negative class data[8, 9], which would increase the probability of false-positive[11], and the generalization ability was not strong. In recent years, some methods add more and more artificial features, even deep learning-based studies[9] ,but feature engineering is time-consuming and laborious and may bring in subjective influence factors. Moreover, most of the past methods used one-hot coding[8, 12], which treated the base sequence as a series of meaningless letters without studying the biological significance of the sequence.

To solve the above problems, a new end-to-end target gene prediction method at the site level is proposed in this paper. From a new perspective, based on the similarity between biological sequences and natural languages, a neural network is used to learn distribution representation of miRNA and mRNA sequences[13]. In our method, miRNA and mRNA were processed as two different languages, and the bases in the sequence were trained as words, referring to the word embedding method commonly used in text-matching tasks. Compared with the local one-hot representation in the past, the distributed representation is used to represent an nt

with a vector, which can be used to describe the similarity distance between nts to a certain extent. In addition, the positive and negative class sample data used in this paper are all verified by a variety of experiments to avoid the problems caused by the mock data in the past. The original sequence data is processed by the word embedding model and directly entered into the neural network for feature extraction and final classification, omitting the manual feature engineering steps. Finally, using 5-fold cross validation, our method has an excellent performance in many aspects on two different data sets, and it is proved superior to two database-based methods and an advanced deep learning method by comparison.

Methods

Datasets

The prediction of miRNA target genes can be divided into site-level and gene-level, and the main difference lies in the different data. The method proposed in this paper deals with site-level prediction, and the data included miRNA sequences and candidate target sites(CTS), and the sample was labeled binding or not. The dataset used in the experiment are all from public databases, and the positive and negative pairs have been verified via biological experiments. We utilize experimental negative data instead of mock ones, and the problem of high false positives in the current prediction model can be solved. The dataset for this experiment consists of two public databases that have been used in recent studies[12]. Diana-Tarbase [14]provided experimentally verified miRNA-mRNA interaction information, including 121,090 positive and 2940 negative pairs. Mirtarbase[15] provides 410,000 positive pairs of miRNA-mRNA interactions. Through screening and deleting data with contradictory results in different experiments and merging duplicate data, the gene-level matching data of 151956 positive and 548 negative pairs verified by many kinds of experiments were finally obtained.

To make a site-level miRNA and CTS pairing information dataset, two databases, PAR-CLIP[16] and CLASH [17], should be utilized to provide the site-level pairing information of miRNA-mRNA verified by experiments. The positive pairs that form stable duplexes, namely, those have negative free energy based on ViennaRNA[18], are remained and complemented by including broadly conserved sites from TargetScanHuman database[19]. Similarly, the negative pairs that have length of up to 30 nts and form stable duplexes are considered as experimentally verified negative pairs. As the result, 33,142 site-level positive and 32,284 site-level negative pairs are used to train the proposed approach.

Distribution representation of miRNA and mRNA sequences

To obtain the distribution and expression of miRNA and mRNA, this article uses the mature miRNA and mRNA sequences of the human genome as corpus, and word2vec[20, 21] is used for training respectively. The process is named mi2vec and m2vec, as shown in Figure 1. The word embedding model is a kind of correlation model that learns the features between words and maps them into dense vectors. This kind of model is a shallow two-layer neural network. In recent bioinformatics studies, some methods[22, 23] have been used to train word embedding models for DNA, proteins, and lncRNAs, and it has been proved that this method is superior

to the traditional processing sequence embedding methods such as one-hot and K-mers.

In this paper, the wor2vec tool in the Gensim package was used for pre-training. The word vector tool was launched by Google in 2013. The training process of the word embedding model was improved, and Negative Sampling and Hierarchical Softmax methods were used to improve the training speed. The training data are derived from miRBase and NCBI databases[24, 25], which store the most authoritative and complete relevant sequence data at present. Both mi2vec and m2vec use a skip-gram network model[26] to obtain the vector expression of bases in the sequence. In word embedding, the embedding dimension is considered to be the most important hyperparameter parameter[27], so vector_ size=2, 4, 20, 50 different output dimensions are set to facilitate the comparison and selection of the optimal parameters in subsequent experiments. The parameters of the model are min_count = 1, window = 5, epoch = 10. Where window stands for maximum distance between the current and predicted word within a sentence, the epoch is the count of iterations over the corpus. When the min_count (means minimum word frequency) is set too high, the model only counts high-frequency words, which is not conducive to learning discriminative word vectors from sequence representation. Other parameters are default.

Table 1 The two layers of RNN selected different architectures for comparison.

First layer	Second layer	accuracy(%)
RNN	RNN	90.93
GRU	GRU	92.96
LSTM	LSTM	93.16
LSTM	BiLSTM	92.87
BiLSTM	LSTM	93.23
BiLSTM	BiLSTM	93.45

BiLSTM

Long short-term memory (LSTM) is an artificial recurrent neural network(RNN) architecture used in the field of deep learning, is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series, and it can deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Because of its design characteristics, LSTM is very suitable for processing text and biological data.

However, it is still impossible to encode information from back to front when using LSTM to model the sequence. BiLSTM, which is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backward direction. BiLSTM effectively increases the amount of information available to the network, improving the content available to the algorithm (e.g. knowing what bases immediately follow and precede a base in a sentence). As shown in Table 1, through experimental comparison of basic RNN, GRU, LSTM, and BiLSTM, the accuracy of LSTM in the model test are higher than that of

other structures, and bidirectional and unidirectional tests are conducted in two-layer stacked LSTM respectively, and it is finally proved that BiLSTM can achieve better results.

Model structure

Figure 2 show an overview of the method proposed in this paper. Firstly, the original sequence data of miRNA and CTS were processed in a uniform length. According to the maximum length of human mature miRNA is 26nts, and the seed region binding to CTS is usually the 2-8 base site[1], so the input sequence was all filled to 30 nts.

Through the vector representation of each base obtained through the mi2vec and m2vec processes in the previous chapter, each base in the input sequence is replaced with a vector, and our input data is finally converted into a 50-dimensional matrix.

In many natural language processing(NLP) studies[28, 29], the method of word embedding combined with RNN has a breakthrough performance. According to the experimental results of different RNN structures in the last chapter, we adopt the BiLSTM for feature extraction. Use BiLSTM to extract the sequence features of miRNA and CTS respectively. Then, to model interactions between miRNA and CTS, the feature maps output by the first layer of BiLSTM are concatenated into one tensor. In this way, two layers of stacked BiLSTM, which have the advantage of learning both the intrinsic spatial and sequential features of miRNA and CTS.

miRNAs target site prediction can treat the target as a dichotomous problem to determine whether miRNA binds or not to CTS, and the sample label indicates whether binding occurs. After the feature extraction via stacked BiLSTM, the feature dimension is gradually reduced to 2-dimensional output by using two linear layers. Finally, the combination relationship was determined by the softmax function.

Results and discussion

In this study, a novel miRNA and CTS interaction prediction model based on sequence distributed representation and deep learning is proposed. In this chapter, the following experiments are designed to verify the performance of the model. Firstly, the effects of one-hot and word2vec coding on the accuracy of deep learning models were compared, and the sequence of distributed representation is visualized and analyzed. Then, a variety of evaluation indexes will be used to verify the performance of the model and compared with other target prediction methods. Finally, each result of this experiment is discussed in depth.

Impact of Distributed Representation on Model

In this paper, BiLSTM and one-dimensional convolutional neural networks(CNN1d), two models which are good at processing sequence data, are selected to test to study the influence of data embedding method on the accuracy of neural network. One-hot coding and four different dimensions of mi2vec and m2vec output were used in the comparison experiment. For training, we optimized the weighted cross-entropy loss function using Adam optimizer [batch size: 32, the number of epochs: 100]. The remaining hyperparameters used were set to be the default PyTorch implementation, and the accuracy on the test set was recorded once every training round.

It can be seen from Figure 3 that the word2vec is used to replace the one-hot on the two network model structures, which effectively improves the accuracy of CNN1d and BiLSTM models. The results show that the method of word embedding is equally effective for biological sequences of miRNA and mRNA. In order to explore the biological significance, the 20-dimensional vector with the best performance on CNN1d was selected as an example for visualization research, and the similarity between bases was analyzed through cosine distance calculation.

It can be seen from the results in the Figure 4 that the distributed representation of sequences has more hidden biological implications. and can better reflect the relationship between bases than the one-hot coding, which can improve the accuracy of the classification model in the deep learning model. As can be seen from the miRNA and mRNA analysis diagram, Figure 4.a and Figure 4.b are visualized heat maps of miRNA and mRNA distributed as 20-dimensional vectors respectively, and it can be seen from the diagram that the distribution of each base is different. By calculating the cosine distance of the vector, it can be seen from Figures 4.c and Figure 4.d that the similarity between the bases in miRNA and mRNA is not the same. In miRNA, the base similarity A and U are the highest, and C and G are the lowest. While in mRNA, the base similarity A and U are the highest, and U and G are the lowest. After one-hot coding, the data presented an orthogonal matrix, and the bases were independent of each other, which could not reflect the information of distributed representation. Therefore, the use of the word2vec can provide more feature information for the deep learning model.

Parameters' effect on the model

The training of the neural network model is determined by its own structural parameters and hyperparameters. In order to obtain the model with optimal performance, the following experiments are designed to determine the model parameters. First, we will compare each layer output unit of BiLSTM and linear layers in order to determine the best model structure. Second, we adjust for the hyperparameters of the training and study the influence of hyperparameters on the experimental results.

Adjusting structural parameters

Table 2 Model structure comparison

First BiLSTM	Second BiLSTM	First Linear	Second Linear	Accuracy(%)
8	32	32	2	92.23
16	32	32	2	92.72
32	32	32	2	93.01
50	32	32	2	93.34
64	32	32	2	92.75
128	32	32	2	92.17
50	8	8	2	92.89
50	16	16	2	92.58
50	64	64	2	92.48
50	128	128	2	92.75

In the study based on RNN parameters, it is found that can better extract sequence features when the input and output dimensions of RNN are the same[30]. Input dimension is represented by a 50-dimension vector of mi2vec and m2vec according to the experimental results of 3.1, so the input dimension is 50, and the

output dimension is also set at 50. In order to verify the previous theory and reference comparison, the first layer BiLSTM set hidden size value [16, 32, 50, 64, 128] for the experimental test, the hyperparameters are temporarily set: batch size =32, lr =0.001, and the optimizer uses Adam.

The adjustment results of the model structure are shown in Table 2. After trying 10 model structures, 50-32-32-2 is finally determined, where the value represents the number of output units of each layer of the network.

Adjusting hyperparameter

Hyperparameters play an important role in the training model. Typical hyperparameters include lr, batch size, the optimizer, etc. We use the usual method of adjusting the hyperparameters: fix all the hyperparameters and then try to modify one of them. Adam has excellent performance is the most widely used in today's deep learning models. In addition, an attempt to use a stochastic gradient descent (SGD) optimizer failed to converge the loss function in this experiment, so Adam is determined to be used as the optimizer. Then adjust lr and batch size respectively, and the results are shown in Table 3.

Table 3 Model hyperparameters comparison

optim	lr	batch_size	Accuracy(%)
Adam	0.0001	32	90.98
	0.0005	32	92.5
	0.001	32	93.34
	0.005	32	91.23
	0.01	32	50.01
	0.001	8	92.11
	0.001	16	92.96
	0.001	64	93.02
	0.001	128	92.53

It can be seen from the results that the model is sensitive to hyperparameters, and some wrong Settings will make the model unable to converge. By comparison, the model achieved the highest accuracy when lr =0.001. Although the selection of batch size has little effect on the results, the larger batch size can significantly reduce the training time. So we made a trade-off and set the batch size =32.

compare different methods

We selected the optimal parameters through a series of experiments mentioned above. In addition, the miRNA and mRNA sequence data were cleaned, and the unverified sequences were removed and the pre-training again. Finally, our model is compared with some site prediction methods.

Evaluation Indicators

As a dichotomous problem of miRNA target gene site prediction, accuracy (Acc), sensitivity (Sens), specificity (Spec), and F-measure are commonly used as evaluation indexes of the comprehensive performance of the model. The calculation formula is as follows:

$$Acc = \frac{TN + TP}{TN + TP + FN + FP}$$

$$Sens = \frac{TP}{TN + FN}$$

$$Spec = \frac{TN}{TN + FP}$$

$$F - measure = \frac{2TP}{2TP + FP + FN}$$

According to the definition of confusion matrix of dichotomy, TP, FP, TN, and FN represent True Positive, False Positive, False Negative, and True Negative respectively.

Performance comparison of two datasets and different methods

In order to prove the generalization ability of our model, the 5-fold cross validation method was used to randomly divide the original dataset into five folds, of which four folds were used as training and one fold as a test. The design experiment was compared with the current commonly used database prediction tools and deep learning target site prediction methods. These include two databases, PITA and TargetScan[31, 32], and two deep learning methods, Lee and DeepmirTar[12, 9]. The DeepmirTar and the data set used in this paper were respectively used for comparative experiments. The results and comparison methods are shown in Table 4.

Table 4 Performance evaluation metrics

Dataset	Method	Acc(%)	Sens(%)	Spec(%)	F-measure(%)
DeepMirTar	TargetScan[a]	58.01	60.23	59.22	NA
	PITA[a]	49.81	58.72	40.82	NA
	DeepMirTar[a]	93.48	92.35	94.79	NA
	Our method[c]	96.86	96.97	96.75	96.91
Our dataset(lee)	TargetScan[b]	55.77	39.45	72.08	47.12
	PITA[b]	50.53	13.65	87.41	21.62
	CNN1d(Lee)[c]	91.05	94.06	87.96	91.40
	Our method[c]	96.04	95.65	96.44	96.09

a: Proposed by DeepMirTar article
b: Proposed by Lee's article
c: 5-fold cross validation results mean
NA:represents the value is not reported

From the comparative experimental results, it can be seen that on the dataset used in the Deepmirtar paper, our method achieves better results than the Deepmirtar method. Moreover, in this paper, the neural network is used to automatically extract features instead of manual feature engineering, which greatly reduces the complexity, saves a lot of time, and has higher accuracy. The negative pairs in the dataset used in this paper have been verified by a variety of experiments to replace the mock data used by DeepMirTar. In comparison with database methods, the data verified by experiments has higher specificity, which can reduce the probability of false-positive in the prediction. Finally, for the experimental data used in this paper, through word2vec pre-training data, combined with the BiLSTM model, and compared with one-hot coding and CNN1d model used in Lee's paper, our method has been improved in Acc Sens Spec and F-measure.

Conclusions

miRNA is an indispensable component of complex transcriptome regulation, which affects life processes and related diseases. To study the function and mechanism of miRNA, the determination of miRNA binding sites is the primary goal. In this study, we developed a deep learning method for predicting miRNA target site by pre-training distribution representation model, mi2vec and m2vec, via using skip-gram word embedding model and human genome-wide miRNA and mRNA sequences. By comparing the performance of different coding methods and other prediction methods, the results prove the effectiveness of the proposed method.

Through other recent literature and our research, it is proved that the NLP method is effective and feasible to deal with the biological sequence problem. The experimental proved this feature extraction scheme works well. However, rethinking of the procedure of mi2vec and m2vec, we recognize that word2vec may not be the best way to mine distribution representation of miRNA and mRNA. How to better understand and express the special language of the biological sequence will become the focus of future research.

Appendix

Text for this section...

Declarations

Abbreviations

mRNAs: messenger RNAs; BiLSTM:Bidirectional Long short-term memory; nts: Nucleotides; RISC: RNA-induced silencing complexes; GRU: Gate recurrent unit; CNN: Convolutional neural network; CTS: Candidate target sites; RNN: Recurrent neural network; NLP: Natural language processing; CNN1d: One-dimensional convolutional neural network; SGD: Stochastic gradient descent; TP: True positives; TN: True negatives; FP: False positives; FN: False negatives.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the Google drive repository, https://drive.google.com/drive/folders/100tZVGFH7jBppI085TblNM9Vuu5gsATQ?usp=sharing .
The pre-training data folder includes the original sequence data of miRNA and mRNA after cleaning, and the vector representation data after word2vec is stored in the mi2VEc and m2vec folders. The training and test data is stored in a CSV file, which is partitioned using PyTorch. The data in the comparative experiment were obtained from the literature[9, 12].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Project of National Natural Science Foundation (61962055 and 31960142)

Authors' contributions

YS(Yuzhuo Sun) conceived the study, collected the datasets and drafted the manuscript. FX provided some ideas and tool support. YS(Yongke Sun),YC supervised the study and provide article guidance. YZ provides guidance on biology-related content. All authors read and approved the final manuscript.

Acknowledgements

Thanks to other authors for their funding and guidance.

Author details

[1]College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, China. [2]College of Material Science and Engineering, Southwest Forestry University, Kunming, China.

References

1. Lu, T.X., Rothenberg, M.E.: Microrna. Journal of Allergy and Clinical Immunology **141**(4), 1202–1207 (2018)
2. Gregory, R.I., Chendrimada, T.P., Cooch, N., Shiekhattar, R.: Human risc couples microrna biogenesis and posttranscriptional gene silencing. Cell **123**(4), 631–640 (2005)
3. Ha, M., Kim, V.N.: Regulation of microrna biogenesis. Nature reviews Molecular cell biology **15**(8), 509–524 (2014)
4. Ruan, K., Fang, X., Ouyang, G.: Micrornas: novel regulators in the hallmarks of human cancer. Cancer letters **285**(2), 116–126 (2009)
5. Rottiers, V., Näär, A.M.: Micrornas in metabolism and metabolic disorders. Nature reviews Molecular cell biology **13**(4), 239–250 (2012)
6. Hsu, S.-D., Tseng, Y.-T., Shrestha, S., Lin, Y.-L., Khaleel, A., Chou, C.-H., Chu, C.-F., Huang, H.-Y., Lin, C.-M., Ho, S.-Y., *et al.*: mirtarbase update 2014: an information resource for experimentally validated mirna-target interactions. Nucleic acids research **42**(D1), 78–85 (2014)
7. Shaker, F., Nikravesh, A., Arezumand, R., Aghaee-Bakhtiari, S.H.: Web-based tools for mirna studies analysis. Computers in biology and medicine, 104060 (2020)
8. Lee, B., Baek, J., Park, S., Yoon, S.: deeptarget: end-to-end learning framework for microrna target prediction using deep recurrent neural networks. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 434–442 (2016)
9. Wen, M., Cong, P., Zhang, Z., Lu, H., Li, T.: Deepmirtar: a deep-learning approach for predicting human mirna targets. Bioinformatics **34**(22), 3781–3787 (2018)
10. Zheng, X., Chen, L., Li, X., Zhang, Y., Xu, S., Huang, X.: Prediction of mirna targets by learning from interaction sequences. Plos one **15**(5), 0232578 (2020)
11. Pinzón, N., Li, B., Martinez, L., Sergeeva, A., Presumey, J., Apparailly, F., Seitz, H.: microrna target prediction programs predict many false positives. Genome research **27**(2), 234–245 (2017)
12. Lee, B.: Deep learning-based microrna target prediction using experimental negative data. IEEE Access **8**, 197908–197916 (2020)
13. Iuchi, H., Matsutani, T., Yamada, K., Iwano, N., Sumi, S., Hosoda, S., Zhao, S., Fukunaga, T., Hamada, M.: Representation learning applications in biological sequence analysis. Computational and Structural Biotechnology Journal **19**, 3198 (2021)
14. Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.-L., Maniou, S., Karathanou, K., Kalfakakou, D., *et al.*: Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. Nucleic acids research **43**(D1), 153–159 (2015)
15. Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J., *et al.*: mirtarbase 2016: updates to the experimentally validated mirna-target interactions database. Nucleic acids research **44**(D1), 239–247 (2016)
16. Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., Rajewsky, N.: Unambiguous identification of mirna: target site interactions by different types of ligation reactions. Molecular cell **54**(6), 1042–1054 (2014)
17. Helwak, A., Kudla, G., Dudnakova, T., Tollervey, D.: Mapping the human mirna interactome by clash reveals frequent noncanonical binding. Cell **153**(3), 654–665 (2013)
18. Lorenz, R., Bernhart, S.H., Zu Siederdissen, C.H., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: Viennarna package 2.0. Algorithms for molecular biology **6**(1), 1–14 (2011)
19. Agarwal, V., Bell, G.W., Nam, J.-W., Bartel, D.P.: Predicting effective microrna target sites in mammalian mrnas. elife **4**, 05005 (2015)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
22. Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., Zhi, D.: Gene2vec: distributed representation of genes based on co-expression. BMC genomics **20**(1), 7–15 (2019)
23. Yi, H.-C., You, Z.-H., Cheng, L., Zhou, X., Jiang, T.-H., Li, X., Wang, Y.-B.: Learning distributed representations of rna and protein sequences and its application for predicting lncrna-protein interactions. Computational and structural biotechnology journal **18**, 20–26 (2020)
24. Griffiths-Jones, S.: mirbase: the microrna sequence database. MicroRNA Protocols, 129–138 (2006)
25. Pruitt, K.D., Tatusova, T., Maglott, D.R.: Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research **35**(suppl_1), 61–65 (2007)
26. McCormick, C.: Word2vec tutorial-the skip-gram model. word2vec-tutorial-the-skip-gram-model (2016)
27. Lai, S., Liu, K., He, S., Zhao, J.: How to generate a good word embedding. IEEE Intelligent Systems **31**(6), 5–14 (2016)
28. Jang, B., Kim, M., Harerimana, G., Kang, S.-u., Kim, J.W.: Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. Applied Sciences **10**(17), 5841 (2020)
29. Muhammad, P.F., Kusumaningrum, R., Wibowo, A.: Sentiment analysis using word2vec and long short-term memory (lstm) for indonesian hotel reviews. Procedia Computer Science **179**, 728–735 (2021)
30. Tan, M., Santos, C.d., Xiang, B., Zhou, B.: Lstm-based deep learning models for non-factoid answer selection. arXiv preprint arXiv:1511.04108 (2015)
31. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., Segal, E.: The role of site accessibility in microrna target recognition. Nature genetics **39**(10), 1278–1284 (2007)
32. Lewis, B.P., Burge, C.B., Bartel, D.P.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. cell **120**(1), 15–20 (2005)

Figure 1 The procedure for training mi2vec and m2vec. Treat each miRNA and mRNA sequence as a special sentence, and treat the bases as the words that make up the sentence. Use the word2vec model to train a vocabulary list composed of bases, and get the vectorized representation of the bases.(The value in the vector in the figure is not the actual result value)

Figure 2 Framework diagram of an end-to-end miRNA target prediction method. Fill the input miRNA and CTS original sequence to a uniform length, and then replace each base in the sequence with the vector trained by mi2vec and m2vec. The processed miRNA and CTS pass through the BiLSTM layer respectively and then concatenate the outputs feature maps, then pass through a layer of BiLSTM, and finally, use the linear layer to reduce the dimension to 2-dimensions.

Figure 3 Different encoding methods and dimension input influence the deep-learning model. (a)Shown is the use of one-hot and word2vec (including training results of 2, 4, 20, and 50 dimensions) as the input of the CNN1d model. (b)The same test using the BiLSTM model shows that word2vec with 50 dimensions as input has the highest accuracy.

Figure 4 Visualization analysis of miRNA and mRNA.(a)The base 20 dimension vector representation of miRNA.(b)The base 20 dimension vector representation of mRNA.(c)The cosine distance between miRNA bases was calculated based on the 20-dimensional vector.(d)The cosine distance between mRNA bases was calculated based on the 20-dimensional vector.

Figures

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.

Figures

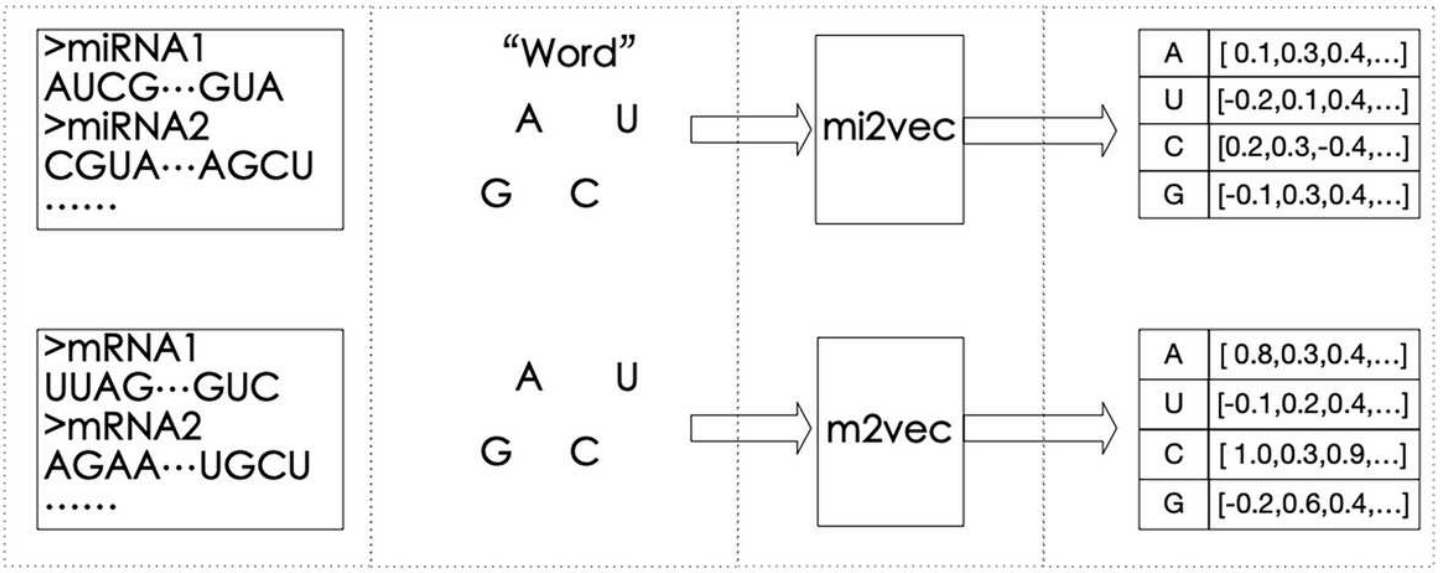

Figure 1

The procedure for training mi2vec and m2vec. Treat each miRNA and mRNA sequence as a special sentence, and treat the bases as the words that make up the sentence. Use the word2vec model to train a vocabulary list composed of bases, and get the vectorized representation of the bases.(The value in the vector in the figure is not the actual result value)

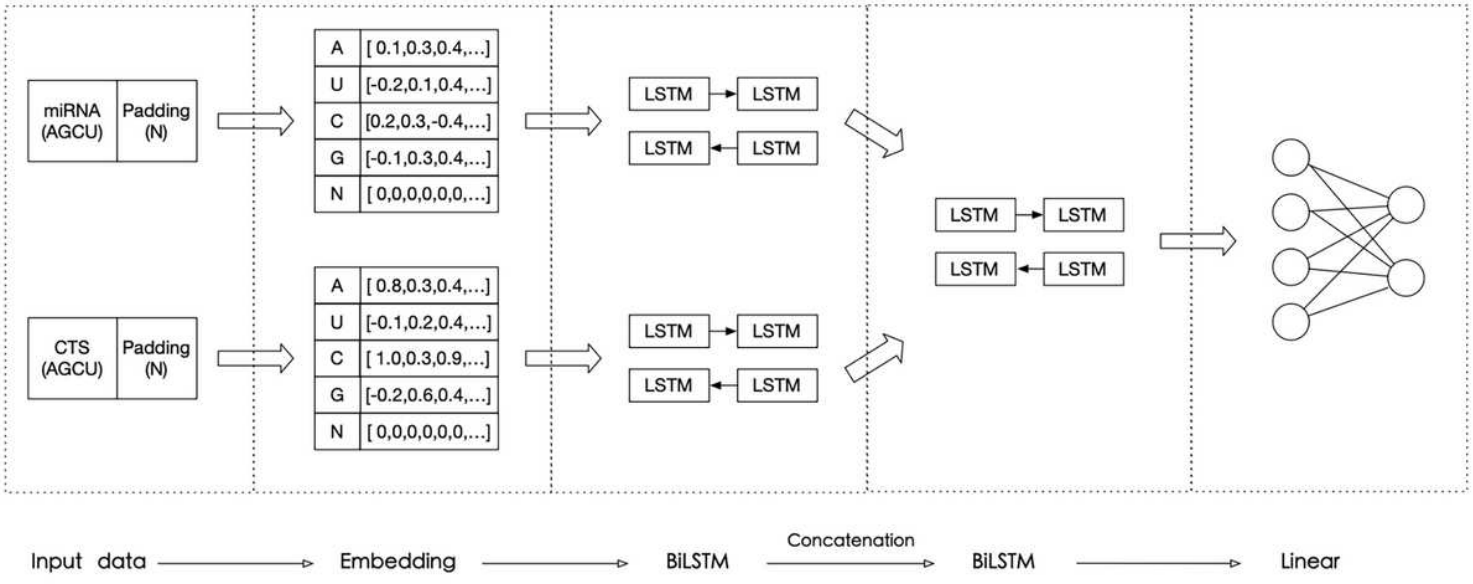

Figure 2

Framework diagram of an end-to-end miRNA target prediction method. Fill the input miRNA and CTS original sequence to a uniform length, and then replace each base in the sequence with the vector trained by mi2vec and m2vec. The processed miRNA and CTS pass through the BiLSTM layer respectively and

then concatenate the outputs feature maps, then pass through a layer of BiLSTM, and finally, use the linear layer to reduce the dimension to 2-dimensions.

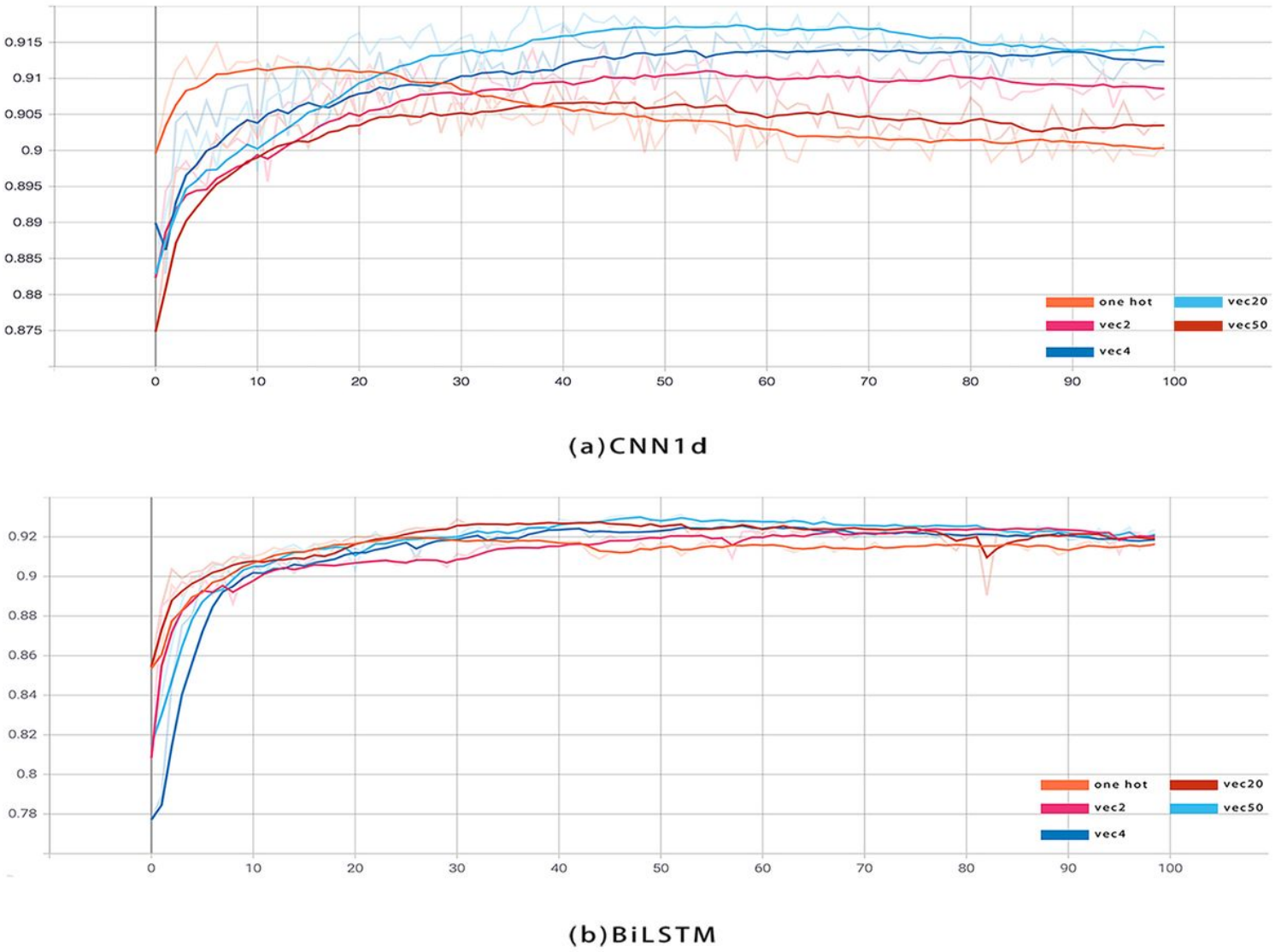

Figure 3

Different encoding methods and dimension input influence the deep-learning model. (a)Shown is the use of one-hot and word2vec (including training results of 2, 4, 20, and 50 dimensions) as the input of the CNN1d model. (b)The same test using the BiLSTM model shows that word2vec with 50 dimensions as input has the highest accuracy.

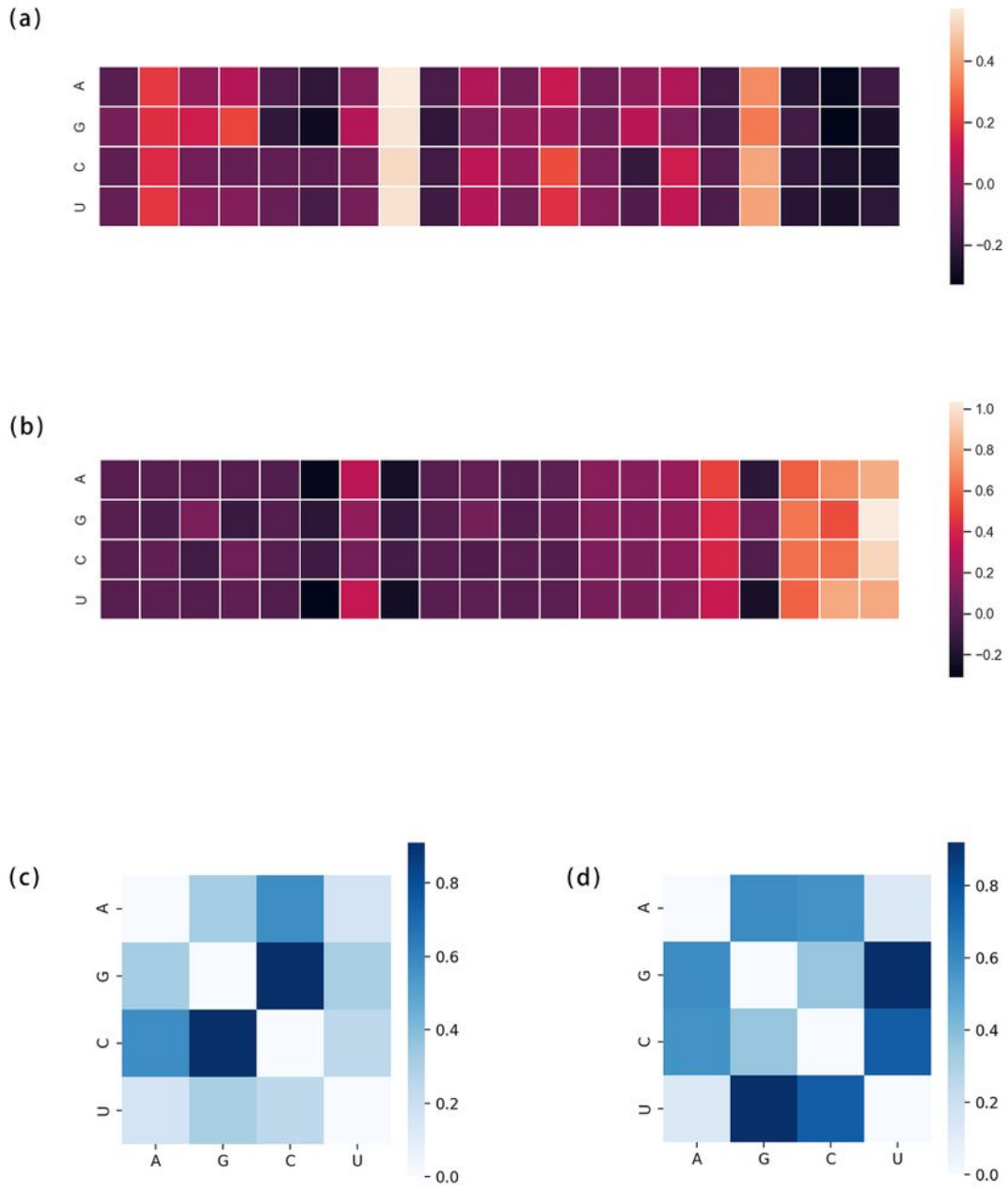

Figure 4

Visualization analysis of miRNA and mRNA.(a)The base 20 dimension vector representation of miRNA. (b)The base 20 dimension vector representation of mRNA.(c)The cosine distance between miRNA bases was calculated based on the 20-dimensional vector.(d)The cosine distance between mRNA bases was calculated based on the 20-dimensional vector.