mirLSTM: A Deep Sequential Approach to MicroRNA Target Binding Site Prediction

Ahmet Paker[1](✉) and Hasan Oğul[1,2](✉)

[1] Department of Computer Engineering, Başkent University,
06790 Ankara, Turkey
pakerahmett@gmail.com

[2] Faculty of Computer Sciences, Østfold University College,
1757 Halden, Norway
hasan.ogul@hiof.no

Abstract. MicroRNAs (miRNAs) are small and non-coding RNAs of ~21–23 base length, which play critical role in gene expression. They bind the target mRNAs in the post-transcriptional level and cause translational inhibition or mRNA cleavage. Quick and effective detection of the binding sites of miRNAs is a major problem in bioinformatics. In this study, a deep learning approach based on Long Short Term Memory (LSTM) is developed with the help of an existing duplex sequence model. Compared with four conventional machine learning methods, the proposed LSTM model performs better in terms of the accuracy (ACC), sensitivity, specificity, AUC (Area under the curve) and F1 score. A web-tool is also developed to identify and display the microRNA target sites effectively and quickly.

Keywords: Deep learning · RNN · LSTM · Bioinformatics · Sequence Alignment · miRNA · Target prediction · miRNA target site

1 Introduction

MicroRNAs (miRNAs) are responsible for biosynthesis and lysis in tissues within cells. They have ~21–23 nucleotides length and play critical and gene-regulatory roles in many living organisms. They bind their partial complementary target site and cause cleavage or posttranscriptional repression. They prohibit the genesis of peptides and output proteins [1]. Recent research shown that gene regulation of psychiatric and neurodevelopmental disorders can be observable because of some miRNAs [2, 3]. Although many miRNA target sites are experimentally and computationally determined, a few numbers of them are experimentally validated. Therefore, computational prediction of the function of miRNA targets is a challenging task to support global effort in understanding gene regulation [12, 13].

The main problem is the elucidation of interaction between microRNAs with their target sites. In this process, interaction is done through biological sequences. Recently, a probabilistic model was created to describe the binding preferences between a microRNA sequence and its target site [4]. They proposed a method which is based on a sequential probabilistic model to express miRNA sequences and its related target site.

© Springer Nature Switzerland AG 2019
G. Anderst-Kotsis et al. (Eds.): DEXA 2019 Workshops, CCIS 1062, pp. 38–44, 2019.
https://doi.org/10.1007/978-3-030-27684-3_6

This model converts miRNA-binding site pair (duplex-sequence) to a new sequence. After that, to analyze a new sequence, they used Variable Length Markov Chain (VLMC) [9].

The classification of sequences is a modeling problem that you have a specific input sequence and predicts the target sequence. The difficulty of this problem is that the sequences can vary in length, consist of a large thesaurus of input characters and that the model should examine the long-term context or dependencies between characters in the input sequence. RNN (Recurrent Neural Network) addresses this issue by adding the feedback mechanism which functions as a memory. Thus, the previous inputs in the model are kept in a kind of memory. LSTM expands this idea by creating a short term and long term memory component. As a result, the LSTM model can give successful results in biological sequences which are made of repeating a set of patterns.

For all these reasons, in this work, five different RNN-LSTM based methods have been proposed to solve the miRNA target site prediction problem. Additionally, miRNA and mRNA binding site model is created which includes probabilistic feature representation approach using miRNA - mRNA pairs. Secondly, LSTM Network Model is created. In detail, firstly we used complementary alignment between miRNA and its relevant binding site. Output alignment is shown as a different sequence. With the help of this sequence, every possible match and mismatch is expressed with different alphabet characters. By using these sequences and using the LSTM model which is described in this study, data were analyzed, trained and classified. Mainly five different methods have been performed and they compared with each other. In the first method named as (Method1), we used small size dataset, which is used in [4] to perform with our deep learning model. The reason for doing Method1 is to improve the work of [4]. Therefore, in addition to the data pre-processing method used in [4] study, we attempted to estimate the target sites of the miRNA by a deep learning method, LSTM. In the second method named as (Method2), we used larger size dataset. In this method, without doing any feature representation methods, raw data in [5] is given directly to our deep learning model to predict miRNA target sites. In the third method named as (Method3), we used the same dataset in Method2, additionally, we used data pre-processing and we intended to set up a classification model which is performed a miRNA target site prediction. In the fourth method named as (Method4), there were used DT (Decision Tree) on the same dataset in (Method2). Lastly, in the fifth method named as (Method5), we used DT on the same dataset in (Method1). In conclusion, we compare the results of five methods with using five metrics which are accuracy (ACC), sensitivity, specificity, AUC (Area under the curve) and F1 score.

Finally, a web server is introduced. This application gets miRNA sequence from the user as a text input and shows its all potential binding sites on the related mRNA sequences with red colour. If LSTM Deep Learning model classifies the output as a “Target” according to the input sequences, outputs (Target Sites) are displayed on the screen. The web server is available at https://mirna.atwebpages.com.

2 Related Work

There exist a variety of tools for miRNA target prediction. DeepMirTar [5] is a recent one based on stacked de-noising auto-encoder deep learning method (SdA) to predict human miRNA-targets on the site level. They used three different feature representations to express miRNA targets. They are: High-level expert designed features, Low-Level expert designed features and Raw-data-level designed features. Seed match, sequence composition, free energy, site accessibility, conservation, and hot-encoding are some of the examples of these features. They achieved 93.48% accuracy, 92.35% sensitivity, 94.79% specificity.

TarPmiR [7] developed a random-forest-based approach to predict miRNA target sites. Their method is based on scanning miRNA on related mRNA sequence to get the perfect seed-matching sites. They used six conventional features and seven of their own features, together. Their method calculates the value of these features. As a result, TarPmiR selects the site which has the highest probability as target-site. They achieved 74.46% accuracy, 73.68% sensitivity, 76.56% specificity.

TargetScan v7.0 [8] states that canonical sites are more functional than non-canonical sites to express miRNA binding sites. They extract 14 new features and train the data with multiple linear regression models. They have reached 58.01% accuracy, 60.23% sensitivity, 59.22% specificity.

3 Materials and Methods

3.1 Embedded Vector Representation

The main concept of the embedded word is that every word used in a language can be represented by a set of numeric values (a vector). Embedded words are N-dimensional vectors that try to capture word-meaning and scope in their values.

Firstly, every duplex sequences obtained by the methodology as discussed in Sect. 3.2 converted letter to index. The character "a" converted index 0, Character "b" converted index 1, Character "c" converted index 2, Character "d" converted index 3, Character "q" converted index 4. The Embedded Word method uses the Euclidean distance to find the relationship between similar sequences. Once the dependencies between characters are found, an embedded vector is obtained (Fig. 1).

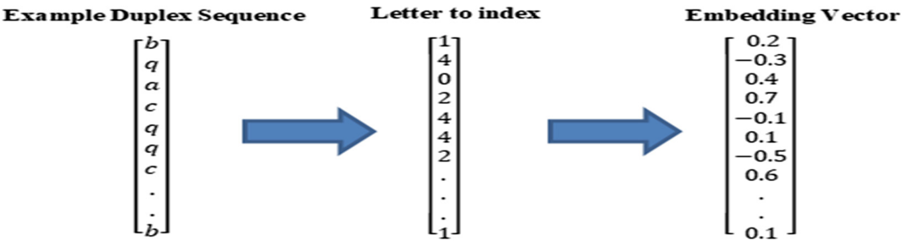

Fig. 1. An example of proposed embedded vector representation

3.2 Proposed LSTM Network

Long short-term memory (LSTM) is a recurrent neural network (RNN) architecture that remembers values at random intervals. Stored values are not changed when learned progress is saved. RNNs allow back and forth connections between neurons.

In our architecture, the first layer is an Embedding layer which will convert our sequences into meaningful numeric embedded vectors as mentioned in Sect. 3.1. In the Embedded layer that uses 32 length vectors to represent each word. The second layer is a Dropout Layer. The main purpose of using this layer is eliminating useless and garbage data. Drop-out percentage is preferred as 20%. The third layer is the LSTM layer with 100 memory units (smart neurons). The fourth layer is a Dropout Layer. Finally, since this is a classification problem, we use a dense output layer with a single neuron and sigmoid activation function to make an estimate of 0 or 1 for two classes in the problem.

Since the binary is a classification problem, log loss is used as a loss function (binary cross-entropy). In addition, as an optimizer "Adam" algorithm is chosen. Adam is an optimization algorithm which updates network weights iteratively in training data. Also to avoid overfitting, in every 5 epoch, validation loss is measured. If validation loss is increased compared to the previous one, early stopping function is activated. As a result, learning is interrupted.

To feed LSTM we use the duplex sequence obtained by complementary alignment of miRNA sequence with a putative target site on mRNA sequence. For every pair of miRNA and target site duplexes are aligned with "Needleman–Wunsch" global complementary alignment algorithm [6]. After that, for every A-U and G-C corresponding nucleotides, different alphabet characters are used. G-U Wobbles are not considered. As a result, it is aimed to decrease the total number of letters in our deep learning model. Each A-U pair is expressed with "a", each U-A pair is expressed with "b", each G-C pair is expressed with "c", each C-G pair is expressed with "d" and all remained pairs are expressed with "q" (Fig. 2).

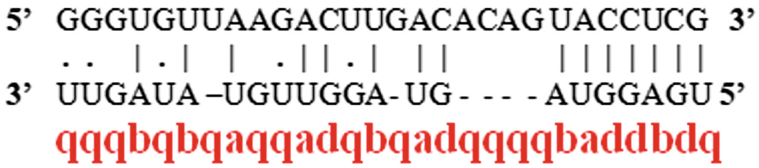
5' GGGUGUUAAGACUUGACACAGUACCUCG 3'
3' UUGAUA–UGUUGGA-UG- - - -AUGGAGU 5'
qqqbqbqaqqadqbqadqqqbaddbdq

Fig. 2. Example alignment of mRNA binding site (top) and miRNA sequence (middle) converted from a new sequence of duplex formation (bottom).

4 Experiments and Results

4.1 Dataset

In this study, two different datasets were used. First dataset is obtained from DeepMirTar repository [5]. In the first dataset, 3915 positive data were obtained, 473 of them are taken from mirMark data [10] and 3442 of them are taken from CLASH data [11]. 3905 negative data were produced using mock miRNAs. Totally there are 7820 data. Since there are lots of experimentally validated positive data, this dataset is preferred. Second dataset is taken from [4]. This dataset contains 283 positive and 115 negative miRNA-mRNA duplex sequences. Totally there are 398 data. The purpose of using this dataset is that the DS2 dataset contains experimentally supported positive and negative duplexes. These two datasets are suitable for comparison because they have different sizes. The size of the first data is larger than the second one. Thus, deep learning method is used for the measurement of working performance of different data with different sizes.

4.2 Empirical Results

Firstly, the raw dataset is randomly mixed. After that, LSTM model was evaluated by determining the test split size as 0.1. Hence, 782 randomly selected test data were gathered. 5 metrics were considered to evaluate success criterion: Accuracy (ACC), sensitivity, specificity, AUC and F1 score. According to Table 1, results show that larger size dataset is more convenient than smaller size dataset for used LSTM model. If Method2 and Method3 are compared, pre-processing (Sequence Alignment Technique as mentioned in Sect. 3) and manipulating data gives more consistent and meaningful results. In conclusion, Method3 is more convenient than Method2. In addition, Method3 gives better results than other four basic machine learning methods (TarPmiR, TargetScan v7.0, Method4, Method5) in terms of evaluation metrics which is discussed on Table 1.

In DeepMirTar, they represent the miRNA-mRNA pairs including 750 features. Some of these features are seed match, free energy, sequence composition, site accessibility, etc. On the other hand, our proposed method represents the miRNA-mRNA pairs is based on a probabilistic approach. In the learning phase, they used SdA (Stacked denoising auto-encoder) based on deep neural network. They split dataset 60% training data, 20% validation data, 20% test data. Besides, we split the dataset 90% training data and 10% test data. They optimized the hyperparameters via grid-search method. On the other hand, we optimized the hyperparameters with random search method. They chose the learning rate of 0.01 and batch size 10. We used the learning rate of 0.1 and batch size 64. Also, they used 1500 memory units (smart neurons) on the other hand we used 100 smart neurons in the LSTM layer.

In conclusion, The DeepMirTar method gave better results than our method, because it has a strong optimization of hyperparameters, a more complex deep neural network structure, and a strong input representation to the first layer of the deep network.

Table 1. Result of evaluation metrics for proposed methods and related works on the same dataset

Methods	Accuracy	Sensitivity	Specificity	F1	AUC
Method1	0.725	0.900	0.200	0.830	0.662
Method2	0.803	0.836	0.770	0.807	0.849
Method3	0.840	0.823	0.856	0.835	0.905
Method4	0.813	0.855	0.825	0.813	0.876
Method5	0.700	0.848	0.150	0.824	0.446
DeepMirTar	0.934	0.923	0.947	0.934	0.979
TargetScan v7.0	0.580	0.602	0.592	0.225	0.672
TarPmiR	0.744	0.736	0.765	0.284	0.802

5 Conclusion

This research revealed an LSTM model in addition to miRNA and mRNA target site duplexes model. In this study, a solution to the problem of how miRNA and mRNA sequences interact and bind is proposed. In addition, the problem of connecting miRNA to the target site of mRNA is solved with the help of the proposed LSTM model.

Classical approaches have implemented via window-based scanner over the mRNA sequence to determine the target binding site. On the other hand, deep learning which is based on the LSTM model gives more consistent and significant results because of input sequences lengths are varied also includes lots of different characters. In addition, the model requires a long-term context or dependencies between each different character in the input sequence. LSTM based deep learning model is quite successful to overcome these problems.

For getting better results, seed and non-seed regions can be considered. The number of layers can be increased by using deep learning model. G-U wobbles can be considered at the pre-processing step. Also the free-energy of miRNA- mRNA duplexes can be conceivable.

References

1. Bartel, D.: MicroRNAs: target recognition and regulatory functions. Cell **136**(2), 215–233 (2009)
2. Bartel, D.: MicroRNAs: genomics, biogenesis, mechanism and function. Cell **116**, 281–297 (2004)
3. Xu, B., Hsu, P., Karayiorgou, M., Gogos, J.: MicroRNA dysregulation in neuropsychiatric disorders and cognitive dysfunction. Neurobiol. Dis. **46**(2), 291–301 (2012)
4. Oğul, H., Umu, S., Tuncel, Y., Akkaya, M.: A probabilistic approach to microRNA-target binding. Biochem. Biophys. Res. Commun. **413**(1), 111–115 (2011)
5. 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)

6. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. **48**, 443–453 (1970)
7. Ding, J., Li, X., Hu, H.: TarPmiR: a new approach for microRNA target site prediction. Bioinformatics **32**, 2768–2775 (2016)
8. Agarwal, V., Bell, G., Nam, J., Bartel, D.: Predicting effective microRNA target sites in mammalian mRNAs. eLife **4**, e05005 (2015)
9. Ron, D., Singer, Y., Tishby, N.: The power of amnesia: learning probabilistic automata with variable memory length. Mach. Learn. **25**, 117–149 (1996)
10. Menor, M., et al.: mirMark: a site-level and UTR-level classifier for miRNA target prediction. Genome Biol. **15**, 500 (2014)
11. Helwak, A., et al.: Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell **153**, 654–665 (2013)
12. Dede, D., Oğul, H.: TriClust: a tool for cross-species analysis of gene regulation. Mol. Inf. **33**(5), 382–387
13. Oğul, H., Akkaya, M.S.: Data integration in functional analysis of microRNAs. Curr. Bioinf. **6**, 462–472 (2011)