Bioinformatics, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Manuscript Category

Subject Section

DeepMirTar: a deep-learning approach for predicting human miRNA targets

Ming Wen[1,*], Peisheng Cong[2], Zhimin Zhang[1], Hongmei Lu[1,*] and Tonghua Li[2,*]

[1]College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China, [2]School of Chemical Science and Engineering, Tongji University, Shanghai 200092, PR, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: MicroRNAs (miRNAs) are small noncoding RNAs that function in RNA silencing and post-transcriptional regulation of gene expression by targeting messenger RNAs (mRNAs). Because the underlying mechanisms associated with miRNA binding to mRNA are not fully understood, a major challenge of miRNA studies involves the identification of miRNA-target sites on mRNA. In silico prediction of miRNA-target sites can expedite costly and time-consuming experimental work by providing the most promising miRNA-target-site candidates.

Results: In this study, we reported the design and implementation of DeepMirTar, a deep-learning-based approach for accurately predicting human miRNA targets at the site level. The predicted miRNA-target sites are those having canonical or non-canonical seed, and features, including high-level expert-designed, low-level expert-designed, and raw-data-level, were used to represent the miRNA-target site. Comparison with other state-of-the-art machine-learning methods and existing miRNA-target-prediction tools indicated that DeepMirTar improved overall predictive performance.

Availability: DeepMirTar is freely available at https://github.com/Bjoux2/DeepMirTar_SdA.

Contact: lith@tongji.edu.cn, hongmeilu@csu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) are endogenous RNAs that play critical regulatory roles, such as cleavage or translational repression, in animals and plants by targeting messenger RNAs (mRNAs) (Bartel, 2004). In mammals, miRNAs are ~23 nucleotides (nts) long (Bartel, 2009), and by combining with argonaute protein, is capable of forming miRNA-induced silencing complex (miRISC) (Thomas, et al., 2010). miRISC can cause mRNA cleavage or translation inhibition, which is guided by miRNA to recognize a specific binding site on mRNA (Thomas, et al., 2010). Traditionally, experimental biochemical assays, such as reporter assays, western blot, quantitative polymerase chain reaction, microarrays, and next-generation sequencing (NGS), can identify miRNA targets(Hsu, et al., 2014). One drawback of these methods is that the targets are identified at the gene level, making it difficult to locate the exact miRNA-binding site on the mRNA(Menor, et al., 2014). Recently, improved cross-linking immunoprecipitation (CLIP) strategies (CLIPL-Seq methods (Wang, 2016)) were applied to experimentally identify miRNA targets at the site level. Helwak et al (Helwak, et al., 2013) developed a crosslinking, ligation, and sequencing of hybrids (CLASH) method to directly observe miRNA-target site, and similarly, Grosswendt et al (Grosswendt, et al., 2014) modified the CLIP methodology to directly and unambiguously identify the miRNA-target site. Besides, In the development of deep sequencing technology that identifying novel miRNAs, lots of individual miRNAs can be found heterogeneous in length and/ or sequence (Morin, et al., 2008). Those variants, which termed as isomiR, refer to those sequences that have variations with respect to the reference miRNA sequence. Recent studies suggest that some isomiRs may affect target selection (Burroughs, et al., 2010; Neilsen, et al., 2012).

The costly and time-consuming experimental identification of miRNA targets promotes reliance on computational tools to predict a set of credible target candidates for further experimental validation. In 2003, hid was found to be targeted by the fly miRNA bantam (Rajewsky, 2006; Stark, et al., 2003), constituting the first case where an miRNA target was identified via bioinformatics. Currently, many in silico approaches have been proposed to predict miRNA targets at the gene and/or site levels (Menor, et al., 2014). These existing methods can be separated into two categories. One is a heuristic method (Fan and Kurgan, 2014), such as Targetscan (Agarwal, et al., 2015; Lewis, et al., 2003), miRanda (John, et al., 2004), PITA (Kertesz, et al., 2007), and PicTar (Krek, et al., 2005), which use screening algorithms to search positions along the miRNA sequence and scoring functions to filter target sites. Early heuristic methods were built based on features, such as complementarity of the miRNA to the target site and/or the free energy of the miRNA/mRNA duplex, which is statistically derived from a limited set of experimental data (Rajewsky, 2006). Due to the increase in data reported from different assays, especially those involving NGS and CLIPL-Seq results, more comprehensive and valuable features have been discovered and implemented in many heuristic methods. TargetScan, first proposed by Bartel et al (Lewis, et al., 2003) in 2003, is one of the most popular heuristic methods and relies upon initial discovery of conserved candidate segments with perfect 2-8 W-C complementary and then filters these segments with folding free energy and other features. In 2015, TargetScan applied sophisticated context+ scores extracted from a relatively large dataset to rank (score) the predicted target sites. The second type of miRNA-target-site prediction involves machine-learning techniques (also called an empirical method (Fan and Kurgan, 2014)), such as mirMark (Menor, et al., 2014), TarPmiR (Ding, et al., 2016), TargetMiner (Bandyopadhyay and Mitra, 2009), and TargetSpy (Sturm, et al., 2010). MiRNA-target-site prediction is a complicated task in bioinformatics that often requires more sophisticated algorithms. Machine-learning methods, such as random forest (RF), support vector machine (SVM) and artificial neural network (ANN) have been frequently used (Ding, et al., 2016; Menor, et al., 2014; Ovando-Vázquez, et al., 2016; Reczko, et al., 2012; Wang, 2016). For a more detailed review of miRNA-target-site prediction, please review (Bottini, et al., 2017; Fan and Kurgan, 2014; Lewis, et al., 2003; Rajewsky, 2006; Reyes and Ficarra, 2012; Riffo-Campos, et al., 2016; Ritchie, et al., 2009; Thomas, et al., 2010; Wagner, et al., 2014; Zheng, et al., 2013)

Recently, deep learning, a branch of machine learning, has become a state-of-the-art performance method applied to various fields of bioinformatics (Alipanahi, et al., 2015; Esteva, et al., 2017; LeCun, et al., 2015; Zhou and Troyanskaya, 2015), including miRNA target prediction (Cheng, et al., 2016). Cheng et al proposed an miRNA-target-site-prediction algorithm called miRTDL, based on a convolutional neural network (CNN). MiRTDL employed the dataset used by TargetScanS(Lewis, et al., 2005) and chose only 20 features to represent miRNA-mRNA pairs, using not the experimentally validated miRNA-mRNA pairs, but rather the miRNA-target duplexes (candidate targets) which met certain feature scores, such as evolutionary conservation score, complementation score, and accessibility score, as the positive sample. It is possible that a positive sample chosen in such an arbitrary way might result in an inaccurate model. The selected 20 features consist of nine complementary features, eight accessible features, and three conservative features. The nine complementary and one of three conservation features are the same as our method, while the remaining accessibility and two of three conservation features are different. In miRTDL, the accessibility and conservation features are represented by regions, such as, the Accessibility_5 (Free energy lost by 5' seed region) and Conservation_1 (Species conversation of target site), while in our method, the accessibility and conservation features are nucleotide-wise, that is, represented by each nucleotide of the miRNA and its targets. In addition, hot-encoding, distance, and more composition features are included in our method.

For miRNA-target recognition, the most commonly used features include seed matching, site conservation, free energy, and site accessibility. A seed match involves a Watson-Crick (W-C) match between an miRNA and its target at the seed region. Depending on the diversity of the seed region, seed matches are classified into several types, such as 6-, 7- and 8-mers (Agarwal, et al., 2015; Helwak, et al., 2013), and is used in almost all miRNA-target-site-prediction methods. Site conservation refers to the commonality of a sequence across species, highly conserved sequences predicted to exhibit biological functions. Through comparative-sequence analyses, Friedman et al (Friedman, et al., 2009) concluded that most mRNA targets represent conserved sequences. Originally, conservation analysis focused on the miRNA-seed region or the mRNA sequence corresponding to the miRNA-seed region. Currently, conservation analysis extends to sequences including both the seed region and its flanking regions (Fujiwara and Yada, 2013; Menor, et al., 2014). Free energy represents a measurement of the stability of a biological system, and is both one of the earliest features employed in miRNA-target-site-prediction tools and a currently indispensable feature in miRNA-target-site prediction. Site accessibility measures the accessibility of mRNA for miRNA binding and considers the contribution of mRNA secondary structure to target recognition (Kertesz, et al., 2007). Previously, site accessibility was not considered by some methods such as miRanda and TargetScan; however, in 2007, Kertesz et al (Kertesz, et al., 2007) experimentally demonstrated that site accessibility plays a critical role in miRNA-target recognition. Currently, site accessibility is a standard feature in miRNA-target-site recognition. In addition, some features were designed following application of machine-learning techniques to miRNA-target-site recognition. These features involve the composition of miRNA/mRNA duplexes and their flanking regions, including the types of target-duplex pairing, the nucleotide (1mer) and dimer (2mer) composition of the region, and local AU content, as well as location of the target site, which indicates the distance of the target from the 3' end of the mRNA.

In this study, an effective deep-learning method – stacked denoising autoencoders (SdA) was applied to accurately predict human miRNA-targets at the site level. Deep-learning methods can learn representations of data with multiple levels of abstraction to discover previously unknown, highly abstract patterns from the nature of the data (raw data) (LeCun, et al., 2015). Although the key characteristic of deep learning is that high-level features are learned from raw data, specialist-designed features are often provided as inputs instead of raw data in many areas of bioinformatics (Min, et al., 2016). One drawback of specialist-designed features is that it may lose certain information of the raw data. To the best of our knowledge, there is no suitable raw-data representational method for miRNA-target prediction; therefore we applied three different levels of features to represent miRNA targets: high-level, expert-designed features, such as seed match, free energy, sequence composition, and target-site location; low-level, expert-designed features, such as site conservation and site accessibility features; and the hot-encoding of sequence data, which is considered a raw-data-level representation. A total of 750 features, including high-level expert-designed, low-level expert-designed, and raw-data-level features were used. We tested our method using test set and compared the results with those obtained from algorithms, such as random forest (RF) (Breiman, 2001), Bernoulli naïve Bayes (BNB) (Metsis, et al., 2006), decision tree (DT) (Anyanwu and Shiva, 2009), logistic regression (LR) (Bishop, 2006), multi-layer perceptron (MLP) (Zhang, 2000) and CNN (Krizhevsky, et al., 2012). Additionally, we compared our method with other existing miRNA-target-site-prediction tools using an external independent dataset with our results indicating that DeepMirTar outperforms other state-of-the-art methods.

2 Methods

2.1. Data

The positive dataset was obtained from two resources: mirMark data (Menor, et al., 2014) and CLASH data (Helwak, et al., 2013). The mirMark data was retrieved from miRecords (Xiao, et al., 2009), and originally contained 507 miRNA-target-site pairs (Additional file 1 (Menor, et al., 2014)). The CLASH data was downloaded from the journal website and contains 18,514 miRNA-target-site pairs (data S2 (Helwak, et al., 2013)). A fully processed mRNA consists of a 5' cap, 5' untranslated region (UTR), a coding sequence (CDS), 3' UTR, and 3' poly-A tail. Although there are target sites located at the 5'UTR and CDS of mRNA, most experimentally validated target sites are located at the 3'UTR (Menor, et al., 2014). Therefore, in this study, only target sites at the 3'UTR would be considered. Additionally, only the canonical seed (exact W-C pairing of 2–7 or 3–8 nts of the miRNA) and the non-canonical seed (pairing at positions 2–7 or 3–8, allowing G-U pairs and up to one bulged or mismatched nucleotide) were considered (Helwak, et al., 2013). All miRNA sequences were retrieved from the mature-miRNA-sequence file (mature.fa) that was downloaded from miRBase (release 21) (Kozomara and Griffiths-Jones, 2014). All 3'UTR sequences were retrieved from the Table Browser of the UCSC Genome Browser (Speir, et al., 2016). Given the possibility that a gene name might correspond to several mRNA accession number, the corresponding mRNA accession number with the longest sequence was used. All miRNA and miRNA-target-site duplexes meeting canonical and non-canonical seed rules were selected using miRanda. Finally, 3915 duplexes were obtained and used as the positive dataset, with 473 duplexes originating from the mirMark data and 3442 duplexes originating from the CLASH data. The list of all positive data is provided in Supplementary File S1.

The negative data were generated using mock miRNAs similar to a previously described method (Menor, et al., 2014). A mock miRNA is obtained by continuously shuffling its corresponding real mature-miRNA sequence until the regions at position two to seven and three to eight in the mock miRNA do not match with any of the same regions of the real mature miRNAs in miRBase (v21). The mock miRNAs were used to screen the corresponding 3'UTR of mRNA using miRanda to randomly select a target site which meeting canonical or non-canonical seed types, which represented one negative sample. Subsequently, 3905 negative samples were generated, with the list of all negative data provided in Supplementary File S2.

Except for the miMark and CLASH dataset, there are no other experimentally validated datasets containing direct miRNA-target site information. An alternative involves the dataset derived from a PAR-CLIP experiment by Hanfer et al (Hafner, et al., 2010), which has been used previously (Ding, et al., 2016; Menor, et al., 2014; Wang, 2016). Although the PAR-CLIP data do not provide exact target-site information, they restrict the target site to a short sequence fragment (~40 nts), which constitute potential positive-target sites and can be considered as an independent test dataset. In practice, if a miRNA is predicted targeting on the fragment, it will be marked as a true prediction. In this study, 48 miRNA canonical- and non-canonical sites which extracted from the PAR-CLIP dataset were set as the independent dataset. The independent dataset consists of 14 miRNAs and 47 mRNAs (Supplementary File S3).

2.2. Features

Seven categories, including 750 features, were chosen to represent the miRNA-target site. The full list of all features and their definitions are provided in Supplementary Information S4. The seven feature categories were further classified into three groups: high-level, expert-designed features, low-level, expert-designed features, and raw-data-level features (Table 1). The calculation methods of each feature categories please see the Features section of Supplementary Information.

Table 1. Feature categories used to represent the miRNA-target site

Category	# of features	Group
Seed match	26	High-level, expert-designed
Free energy	5	
Sequence composition	98	
Site location	1	
Site conservation	160	Low-level, expert-designed
Site accessibility	370	
Hot-encoding	90	Raw-data-level

2.3. Stacked denoising auto-encoder (SdA)

An auto-encoder is an unsupervised graphic model that learns the representation from input data (Bengio, 2009) and is comprised of a two-layered network consisting of visible (input) and hidden layers (Fig. 1A). The auto-encoder is trained to encode the input, v, into some representation, h, so that the input to be reconstructed from that representation, with the training of the autoencoder involving an encoding step (encoder) and a decoding step (decoder). As shown in Fig. 1B, in the encoding step, the input, v, is mapped to a hidden representation, h, through deterministic mapping:

$$h = s\left(Wv + b\right) \tag{1}$$

where s is a non-linear function, such as the sigmoid function, W represents the weight that connects the visible layer and the hidden layer units, and b is the offset of the visible layer. In the decoding step, the latent representation, h, is then mapped back into a reconstruction, v', of the same shape as v through a similar mapping transformation:

$$v' = s\left(W'h + b'\right) \tag{2}$$

The reconstructed $v^{'}$ represents a prediction of v. The parameter $\theta = \left\{W, W^{'}, b, b^{'}\right\}$ of the autoencoder model is optimized by minimizing the average reconstruction error. The reconstruction error, $L(v, v')$, can be measured as:

$$L\left(v, v'\right) = -\sum_{k=1}^{d}\left[v_k \log v_k^{'} + \left(1 - v_k\right)\log\left(1 - v_k^{'}\right)\right] \tag{3}$$

where d is the length of v.

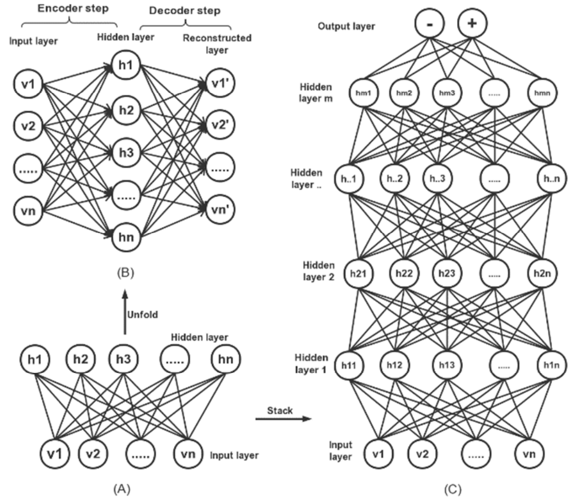

Fig. 1. Autoencoder and SdA

A dA is the same as an autoencoder, except that stochastic noise is added to the input layer (Fig. 2B) (Vincent, et al., 2008) . In practice, this is performed by randomly forcing some units of the visible layer to be zero, while the others are left unchanged (Vincent, et al., 2008).

$$v_{denoised} = r\left(v\right) \tag{4}$$

$$h = s\left(Wv_{denoised} + b\right) \tag{5}$$

while r is a function that used to randomly force some units of input to be zero. h is then encoded from $v_{denoised}$ and the decoder step and reconstruction error of dA are the same as autoencoder. The motivation for the denoising autoencoder is to discover robust representations. For more information about applying denoising procedure in autoencoder please see (Vincent, et al., 2008). The dA is an extension of a classical autoencoder that was introduced as a building block for deep networks (Vincent, et al., 2008). As shown in Fig. 1C, dAs are stacked to form a deep network, resulting in a SdA. Except for the top output layer, every two adjacent layers form a dA, with the representation layer of the dA serving as the input layer of the current dA. The representation layer of the last dA and the top output layer form a logistic regression network. The unsupervised pre-training of an SdA is performed on one dA at a time, with each dA trained by minimizing the reconstruction error. Once all dAs are pre-trained, the entire network, including the top output layer, is trained to minimize the negative log-likelihood loss:

$$l\left(\theta, D\right) = -\sum_{i=0}^{|D|}\log\left(P\left(Y = y^{(i)} \mid v^{(i)}, \theta\right)\right) \tag{6}$$

where $\theta = (W_1, b_1, W_2, b_2, \cdots, W_{m+1}, b_{m+1},)$, m is the number of hidden layers. D is the training data that used to train the model. Y is the corresponding true label of D. This procedure is called supervised fine-tuning that aims to minimize prediction error on a supervised task.

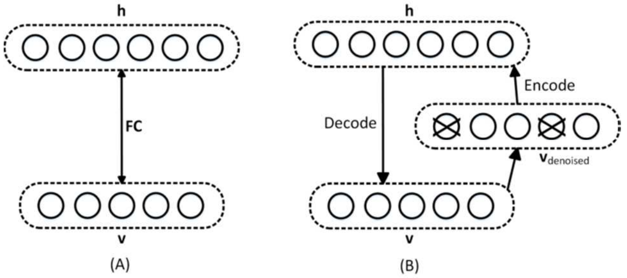

Fig. 2. Denoising autoencoder. (A) a basic autoencoder, FC means full-connected, (B) a denoising autoencoder, the input v is corrupted to $v_{denoised}$, the encoder then encodes $v_{denoised}$ to h and the decoder attempts to reconstruct v.

3 Results

3.1 Determination of the SdA architecture and hyperparameters

The training of the SdA requires the determination of multiple of hyperparameters. An SdA is a deep neural network that consists of multiple layers, with massive units on each layer (Fig. 1C). SdA has several hyperparameters, including the number of hidden layers (nhl), the number of units in each layer (nul), a pre-training rate (pr), a fine-tuning rate (fr), a pre-processing method (pm), and the batch size (bs). These hyperparameters were determined by minimizing the prediction accuracy on the validation dataset through a two-step grid-search approach. The first step involves a grid search using large grid spacing, followed by a fine-tuning grid search with small grid spacing (Table S2). Other hyperparameters, such as the training and pre-training epochs, were set to 120 and 800, respectively. Over 800 models were built to optimize the parameters, with the optimized pr, fr, bs and pm at 0.2, 0.005, 1 and MinMaxScaler, respectively. The SdA architecture resulted in [750, 1500, 2000, 2000, 2000, 1500, 2], where '750' represented the number of units of the visible layer and '2' the number of units of the output layer, with the other numbers representing the number of units in the hidden layers.

3.2 Comparison of different machine-learning methods using a test dataset

To evaluate the performance of different machine-learning methods on miRNA-target prediction, a dataset was split into three subsets consisting of a training set (60% of the original dataset), a validation set (20%), and a test set (20%), with the training set used to train the classifier, the validation set used to optimize the hyperparameters, and the test set used to evaluate the performance of each classifier. Seven commonly used machine-learning methods, including DT, BNB, LR, RF, MLP, CNN and SdA, were considered. We did not consider SVM due to the similar performance observed between RF and SVM in bioinformatics fields, as well as the time-saving nature of using RF to train the model. The performance results for the different methods are listed in Table 2. We found that SdA outperformed DT, BNB, LR, RF, CNN and MLP in AUC, ACC, TPR, and TNP ($p < 0.0001$; Student's t-test). Compared with RF, which is the most frequently used machine-learning method for miRNA-target-site prediction, SdA results in increases in AUC, ACC, TPR, and TNR values of 3.11%, 5.74%, 4.6% and 6.6%, respectively. MLP has a similar architecture to that of the SdA, with both consisting of multiple layers and multiple units in each layer. The results obtained from using either an MLP or the SdA indicated that each performed better than the other methods, indicating that deep neural network exhibited improved ability to learn. Additionally, compared with MLP, the SdA exhibited an improved performance indicating that unsupervised pre-training is beneficial to the deep-neural-network model. Overall, the SdA exhibited the best performance in TPR, TNR, ACC, and AUC among all models, indicating that the SdA model was reliable and can be further applied for novel miRNA-target prediction.

Table 2. The overall performance of different machine-learning methods on the test dataset.

Methods	Testing data set						p-value [a]
	AUC	ACC	TPR	TNR	F score	MCC	
DT [b]	0.8768 (0.0127) [c]	0.8139 (0.0137)	0.8558 (0.0258)	0.8259 (0.0279)	0.8139 (0.0136)	0.6313 (0.0277)	$p < 10^{-6}$
BNB [b]	0.8583 (0.0082)	0.7570 (0.0098)	0.8626 (0.0176)	0.7039 (0.0222)	0.7570 (0.0098)	0.5252 (0.0196)	$p < 10^{-24}$
LR [b]	0.9244 (0.0069)	0.8491 (0.0117)	0.8317 (0.0207)	0.8677 (0.0244)	0.8370 (0.0075)	0.6773 (0.0133)	$p < 10^{-27}$
RF [b]	0.9489 (0.0067)	0.8811 (0.0090)	0.8774 (0.1105)	0.8851 (0.0110)	0.8817 (0.0089)	0.7634 (0.0177)	$p < 10^{-14}$
MLP[b]	0.9568 (0.0058)	0.8990 (0.0099)	0.9044 (0.1192)	0.8934 (0.0131)	0.8765 (0.0093)	0.7528 (0.0170)	$p < 10^{-4}$
CNN-1D [b]	0.9505 (0.0220)	0.8886 (0.0145)	0.8735 (0.0211)	0.8821 (0.0098)	0.8848 (0.0114)	0.7686 (0.0218)	$p < 10^{-4}$
CNN-2D [b]	0.9410 (0.0198)	0.8765 (0.0169)	0.8701 (0.0150)	0.8892 (0.0073)	0.8810 (0.0082)	0.7623 (0.0164)	$p < 10^{-4}$
SdA	0.9793 (0.0100)	0.9348 (0.0205)	0.9235 (0.0468)	0.9479 (0.0180)	0.9348 (0.0371)	0.8699 (0.0205)	----

a The p-value according to Student t-test (ACC) and indicating comparison between the SdA and other methods.

b The optimized Hyperparameters please see Supplementary Information.

c The number in the bracket indicates the standard deviation of results from 20 models built based on different random split of training, validation and test dataset.

3.3 Comparison of DeepMirTar with existing miRNA-target prediction methods

The test dataset, which was randomly separated from the original dataset with a ratio of 0.2, was used to compare the performance of SdA with existing available miRNA-target-prediction tools. Five well-known miRNA target prediction tools, including miRanda, RNAhybrid, PITA, TargetScan v7.0, and TarPmiR, were selected and tested (Supplementary Information). Because the definition of the target site and the location of the target site might vary between different methods, a CTS was considered as predicted 'true targets site' when a predicted target site was located in the miRNA-target-site sequence. Table 3 shows the overall performance results of different methods using this dataset. Our results indicated that DeepMirTar outperformed all methods according to AUC, ACC, TPR, and TNR. To further evaluate the DeepMirTar performance, we evaluated it on an independent dataset. The performance of different tools was listed in Table 3. In agreement with results of the previous dataset, DeepMirTar outperformed the other tools, indicating that this method was the most effective at miRNA-target prediction between two different test datasets.

Table 3. The overall performance of different existing methods on the test dataset and the independent test dataset.

Methods	Test data set				Independent set	Latest version
	AUC	ACC	TPR	TNR	TPR	
miranda	0.6874	0.6592	0.6522	0.6662	6/48	2010
RNAhybrid	0.7585	0.6988	0.6446	0.7535	3/48	2010
PITA	—	0.4981	0.5872	0.4082	23/48	2015
TargetScan v7.0 [a]	0.6725	0.5801	0.6023	0.5922	17/48	2015
TarPmiR	0.8021	0.7446	0.7368	0.7656	19/48	2016
DeepMirTar (SdA)	0.9793	0.9348	0.9235	0.9479	24/48	—

[a] Instead of using the mock RNAs which do not have miRNA family name, the negative pairs in miRTDL were used to evaluate the TargetScan.

3.4 Feature importance evaluated by RF

RF represents an ensemble of simple tree predictors, where each tree is built based on samples that are randomly selected from the original samples. Similarly, features used to represent a sample are also randomly selected from the original features. The importance of the features can be ranked by feature-importance score reported by RF (Breiman, 2001). Because the SdA model is a black box, it is difficult to obtain insight into the features through this method. Here, the importance scores associated with the features and evaluated by RF were employed to investigate the importance of the features. Table S5 lists 22 features with importance score >5 (all importance scores are listed in Supplementary File S5). Among the 22 features, some were previously reported as being important for miRNA-target prediction, including Total_GC, Total_bulge_nt, Total_mismatch and Total_AU in the top 10 selected features (Menor, et al., 2014; Wang, 2016). Additionally, Seed_match_7mer2, Seed_match_6mer3GU1, and Seed_match_6mer2 are seed match features frequently used for miRNA-target prediction, and MEF_Duplex, a free energy feature, is also commonly used for these methods. Because these features are used in other methods, their high-performance scores were expected; however, 13 hot-encoding features, also exhibited high importance scores. Hot-encoding features involve the digital representation of each base in the target sequence, with each base of the sequence represented by a vector of length five. Among all the hot-encoding features, our results indicated that positions 9, 8, 7, 6, 5, 2, and 1 of miRNA and position 5 of mRNA were more important than the other positions. Fig. 3 represents a plot of feature-importance scores for each category. The seed-match and sequence-composition categories were merged into a composition category because they represent features related to the composition of miRNAs and their targets. We observed that some features associated with the composition and hot-encoding categories were more important than other features according to the RF model. The minimum free energy features had higher importance scores as compared with site accessibility and site conservation features, but lower importance scores relative to some composition and hot-encoding features. An interesting phenomenon involved dramatic changes in importance score in the composition and hot-encoding features. Although the importance scores of sites accessibility was not very large, changes in the scores were smooth, with most of the features exhibiting importance scores >0.5, indicating the accessibility category play a role in RF classifier. This might be explained by the dependence of the effect of conservation and accessibility on the entire target-site sequence. For example, site accessibility is a measurement of the accessibility of the mRNA to miRNA binding, which includes the requirement that all bases of the target be accessible to the miRNA. Therefore, all the bases have a similar contribution to miRNA binding. Overall, several hot-encoding, composition, and free energy features have significant contributions to the RF model, while the conservation, and accessibility categories also play a role in RF classifier.

According to the RF feature importance score, we test the machine learning methods with the top L % features (L is 10, 30, and 50). The results were listed in Table S6 in the Supplementary Information. From Table S6 we can find that there is no distinct improvement for DT, BNB, and LR with the increasing of features, which infers that the top-ranked features are significantly important and the remaining features do not have 'antagonistic effect' on these methods. However, for RF, the increasing of features resulting in worse performances. This may come from the 'antagonistic effect' between features. For these deep neural network method (MLP, CNN-1D, CNN-2D, and SdA), the performance fluctuated a lot and there is no obvious change rule when using 10%, 30%, 50%, and 100% features.

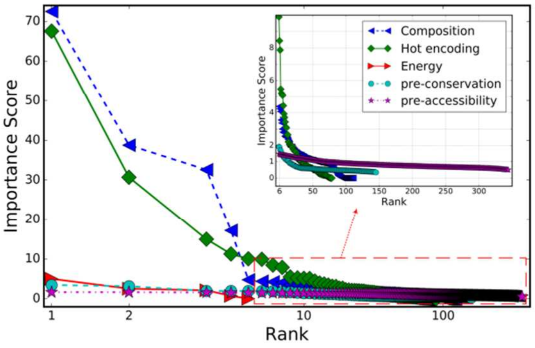

Fig. 3. Feature importance score of each category. The horizontal axis is the rank of feature importance score in a category. The vertical axis is the importance score of a feature.

3.5 The influence of unsupervised pre-training

MLP and SdA achieved the best performance among all methods; however, the difference between MLP and SdA cannot be neglected. MLP and SdA have similar architectures, which include tremendous weights and biases. The initial weights and biases are set artificially, and a stochastic gradient descent (SGD) strategy is often used for their optimization. The initial weights and biases have an important effect on optimization results, and instead of optimizing based on the initial assignment, as with MLP, the SdA applies a layer-wise, unsupervised pre-training procedure (auto-encoder) to 'pre-learn' the weights and biases from the initial assignment and then employs the pre-learned weighted and biases to perform SGD. The 'pre-learned' weights and biases might enable the SDG to determine a better terminal optimization point. Moreover, unsupervised learning along with supervised learning is particularly beneficial to miRNA-target prediction. In the unsupervised procedure, data consist of samples that do not require labels; therefore, all the CTS can be used to feed to the pre-training procedure. With more data used in pre-training procedure, the distribution of the training dataset is much closer to the actual distribution of the miRNA-target site, resulting in decreased model bias.

The success of a machine learning algorithm generally depends upon the features used to represent and describe the data. This is because different features can entangle and hide different explanatory factors associated with variations behind the data, which encourages experts to design more powerful features to help algorithms perform better on a certain task. The importance of features highlights the weakness of traditional learning algorithms: they are unable to extract and organize discriminative information from the data. Experts design more powerful features as a way of compensating for algorithmic weaknesses by taking advantage of human ingenuity and prior knowledge. With the goal of yielding more abstract and useful representations, the SdA can yield abstract and useful representations based on the weights and biases, with the hidden layers of SdA formed by the composition of multiple non-linear transformations of the input data using the weights and biases. The weights and biases originate from initialization, pre-training, and training. Except for initialization, the last two steps attempt to abstract the original representation to a type of high-level representation that will benefit the final prediction of the miRNA-target site.

3.6 CONCLUSION

Because many diseases are related to miRNA activity, identifying miRNA-target sites is important to the process of revealing their biological function. Computational predictions of potential miRNA-targets sites, followed by experimental validation, increase the speed of miRNA-target site identification and reduce the money; however, the lack of reliable experimental data for use in building prediction models remains an obstacle. Furthermore, miRNA-target prediction represents an intricate classification problem that requires sophisticated machine-learning methods to build the prediction models. In this study, we combined deep-learning method with a relatively large dataset to develop an miRNA-target-prediction method called DeepMirTar capable of considering 750 features from different levels. Results on multiple test dataset showed that this method outperformed current target-prediction tools.

Funding

This work has been supported by the National Natural Science Foundation of China [21375151, 21675174, and 21305163 to Hongmei Lu, 21275108 to Tonghua Li] and the Hunan Provincial Innovation Foundation for Postgraduate (CX2017B044).

Conflict of Interest: none declared.

References

Agarwal, V., *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.

Alipanahi, B., *et al.* (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, **33**, 831-838.

Anyanwu, M.N. and Shiva, S.G. (2009) Comparative analysis of serial decision tree classification algorithms. *International Journal of Computer Science and Security*, **3**, 230-240.

Bandyopadhyay, S. and Mitra, R. (2009) TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*, **25**, 2625-2631.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, **116**, 281-297.

Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *cell*, **136**, 215-233.

Bengio, Y. (2009) Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, **2**, 1-127.

Bishop, C.M. (2006) Pattern recognition. *Machine Learning*, **128**, 1-58.

Bottini, S., *et al.* (2017) Recent computational developments on CLIP-seq data analysis and microRNA targeting implications. *Briefings in Bioinformatics*.

Breiman, L. (2001) Random forests. *Machine learning*, **45**, 5-32.

Burroughs, A.M., *et al.* (2010) A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome research*, **20**, 1398-1410.

Cheng, S., *et al.* (2016) MiRTDL: a deep learning approach for miRNA target prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, **13**, 1161-1169.

Ding, J., *et al.* (2016) TarPmiR: a new approach for microRNA target site prediction. *Bioinformatics*, **32**, 2768-2775.

Esteva, A., *et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, **542**, 115-118.

Fan, X. and Kurgan, L. (2014) Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Briefings in bioinformatics*, bbu044.

Friedman, R.C., *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, **19**, 92-105.

Fujiwara, T. and Yada, T. (2013) miRNA-target prediction based on transcriptional regulation. *BMC genomics*, **14**, S3.

Grosswendt, S., *et al.* (2014) Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Molecular cell*, **54**, 1042-1054.

Hafner, M., *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129-141.

Helwak, A., *et al.* (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654-665.

Hsu, S.-D., *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic acids research*, **42**, D78-D85.

John, B., *et al.* (2004) Human microRNA targets. *PLoS Biol*, **2**, e363.

Kertesz, M., *et al.* (2007) The role of site accessibility in microRNA target recognition. *Nature genetics*, **39**, 1278-1284.

Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, **42**, D68-D73.

Krek, A., *et al.* (2005) Combinatorial microRNA target predictions. *Nature genetics*, **37**, 495-500.

Krizhevsky, A., *et al.* Imagenet classification with deep convolutional neural networks. In, *Advances in neural information processing systems*. 2012. p. 1097-1105.

LeCun, Y., *et al.* (2015) Deep learning. *Nature*, **521**, 436-444.

Lewis, B.P., *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell*, **120**, 15-20.

Lewis, B.P., *et al.* (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.

Menor, M., *et al.* (2014) mirMark: a site-level and UTR-level classifier for miRNA target prediction. *Genome biology*, **15**, 500.

Metsis, V., *et al.* Spam filtering with naive bayes-which naive bayes? In, *CEAS*. 2006. p. 28-69.

Min, S., *et al.* (2016) Deep learning in bioinformatics. *Briefings in Bioinformatics*, bbw068.

Morin, R.D., *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome research*, **18**, 610-621.

Neilsen, C.T., *et al.* (2012) IsomiRs–the overlooked repertoire in the dynamic microRNAome. *Trends in Genetics*, **28**, 544-549.

Ovando-Vázquez, C., *et al.* (2016) Improving microRNA target prediction with gene expression profiles. *BMC genomics*, **17**, 364.

Rajewsky, N. (2006) microRNA target predictions in animals. *Nature genetics*, **38**, S8-S13.

Reczko, M., *et al.* (2012) Accurate microRNA target prediction using detailed binding site accessibility and machine learning on proteomics data. *Frontiers in genetics*, **2**, 103.

Reyes, P.H. and Ficarra, E. (2012) One decade of development and evolution of microRNA target prediction algorithms. *Genomics, proteomics & bioinformatics*, **10**, 254-263.

Riffo-Campos, Á.L., *et al.* (2016) Tools for sequence-based miRNA target prediction: What to choose? *International journal of molecular sciences*, **17**, 1987.

Ritchie, W., *et al.* (2009) Predicting microRNA targets and functions: traps for the unwary. *Nature methods*, **6**, 397-398.

Speir, M.L., *et al.* (2016) The UCSC genome browser database: 2016 update. *Nucleic acids research*, **44**, D717-D725.

Stark, A., *et al.* (2003) Identification of Drosophila microRNA targets. *PLoS Biol*, **1**, e60.

Sturm, M., *et al.* (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC bioinformatics*, **11**, 292.

Thomas, M., *et al.* (2010) Desperately seeking microRNA targets. *Nature structural & molecular biology*, **17**, 1169-1174.

Vincent, P., *et al.* Extracting and composing robust features with denoising autoencoders. In, *Proceedings of the 25th international conference on Machine learning*. ACM; 2008. p. 1096-1103.

Wagner, M., *et al.* (2014) MicroRNA target prediction: theory and practice. *Molecular Genetics & Genomics*, **289**.

Wang, X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, **32**, 1316-1322.

Xiao, F., *et al.* (2009) miRecords: an integrated resource for microRNA–target interactions. *Nucleic acids research*, **37**, D105-D110.

Zhang, G.P. (2000) Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **30**, 451-462.

Zheng, H., *et al.* (2013) Advances in the Techniques for the Prediction of microRNA Targets. *International journal of molecular sciences*, **14**, 8179-8187.

Zhou, J. and Troyanskaya, O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, **12**, 931-934.