Effective small interfering RNA design based on convolutional neural network

1st *Ye Han*
School of Information Technology
Jilin Agricultural University
Changchun, China
hanye311@126.com

2nd *Fei He*
School of Information Science and Technology
Institute of Computational Biology
Northeast Normal University
Changchun, China
hef740@nenu.edu.cn

3rd *Xian Tan*
School of Information Science and Technology
Institute of Computational Biology
Northeast Normal University
Changchun, China
tanx431@nenu.edu.cn

4th *Helong Yu*
School of Information Technology
Jilin Agricultural University
Changchun, China
264496469@qq.com

***Abstract*—In functional genomics, small interfering RNA (siRNA) can be used to knockdown gene expression. Usually, a target gene has numerous potential siRNAs, but their efficiencies of gene silencing often varies. Thus, for a successful RNA interference (RNAi), selecting the most effective siRNA is a critical step. Despite various computational algorithms have been developed, the efficacy prediction accuracy is not so satisfactory. In this paper, to explore the effect of different motifs on gene silencing and further improve the prediction accuracy, we developed a new powerful predictor by using a deep learning algorithm—Convolutional Neural Network (CNN). The comparison results showed that the Pearson Correlation Coefficient (PCC) of our model is 0.717, which is 13.81%, 16.78% and 5.91% higher than Biopredsi, i-Score, ThermoComposition21 and DSIR. In addition, the area under the ROC curve (AUC) of our model is 0.894, which is 10.10%, 12.59% and 7.07% higher than those four algorithms. The results show that our model is stable and efficient to predict siRNA silencing efficacy.**

Keywords- siRNA; deep learning; Convolutional Neural Network; RNAi

I. INTRODUCTION

RNA interference (RNAi) is the process that small interfering RNA (siRNA) induces sequence specific post-transcriptional gene silencing [1-3]. RNAi was discovered in many eukaryotic systems, including plants, fungi, invertebrates and mammals [4]. In mammalian cells, double-stranded RNA (dsRNA) is processed into short 21–23 nucleotide (nt) siRNA and induces instant target gene knockdown [3]. To date, siRNA has become an important tool for studying gene function [5-7] and provided a new technology for the treatment of influenza virus [8], HIV virus [9], and even cancer [10]. Compared with other gene tools, siRNA-based gene silencing is becoming more and more popular for its simplicity and low cost.

In the process of RNAi, siRNA design is one of the most critical steps and many efforts are being made in this area. Numerous efficacy prediction methods based on machine learning algorithms were developed. Huesken [11] developed a siRNA efficacy prediction tool 'Biopredsi' based on artificial neural network and built a major siRNA dataset including 2431 siRNAs through high-throughput analysis technology, which created a new era in siRNA efficacy prediction area. Many siRNA efficacy prediction algorithms were built using Huesken's data set[12]. ThermoComposition-21[13] is also an artificial neural network algorithm, which combines both thermodynamic and composition features. To design an accurate and easily interpretable efficacy prediction model, Jean-Philippe Vert[14] proposed a simple linear method combining two simple sets of siRNA sequence features, including the nucleotides present at each position in the siRNA sequence and the global content of the siRNA in short motifs. The efficacy prediction accuracy of the simple linear model is as accurate as the Biopredsi neural network. i-score[15] is a simple siRNA efficacy prediction algorithm based on a linear regression model too. The algorithm is exclusively composed of the nucleotide preference scores and can predict active siRNAs to the similar extents as Biopredsi, ThermoComposition-21 and DSIR.

Despite the considerable efforts, the efficacy prediction accuracy is not so satisfactory. Most machine learning algorithms highly depend on the siRNA features, including nucleotide preference, nucleotide composition, thermodynamic stability, secondary structure and so on. And the prediction ability is limited by these features.

To further improve the prediction accuracy and explore the effect of different motifs on gene silencing, we developed a new powerful predictor based on Convolutional Neural Network (CNN)[16, 17]. Unlike previous efficacy prediction methods, our model utilized CNN algorithm to automatically learn motif encoding features. In the convolution layer of our CNN model, the convolution kernels were designed as motif detector to automatically learn the potential feature pattern of siRNA multimode motif in a data-driven method, which is more abstract, more closely to the essence and more conducive to classification. Compared with the current

siRNA efficacy prediction methods, our model performed best in terms of prediction accuracy.

II. METHODS

A. Our CNN model

In this paper, we developed a siRNA silencing efficacy prediction model based on convolutional neural network. Combining sample size and computational complexity, we defined three significant layers, including convolutional layer, pooling layer and output layer. The structure of our model is shown in Figure 1.

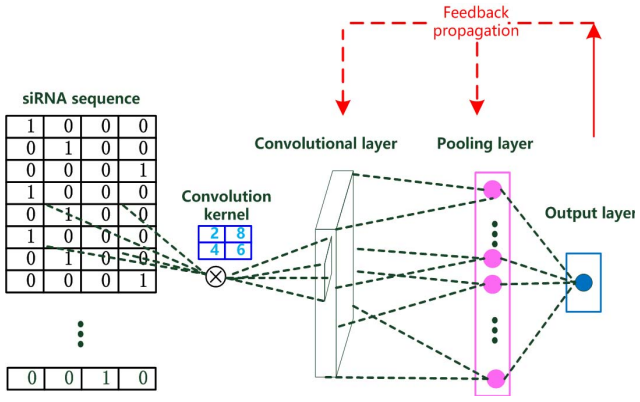

Figure 1. The structure of our CNN model

In the convolutional layer, multiple convolution kernels with different size are designed as motif detector to seek the motifs which are significant to siRNA silencing efficacy. The convolution kernel can be considered as the weights extracted from the position and composition of siRNA multimode motif. And the convolution result is the corresponding motif encoding features. Different from the existing sequence encoding rules defined by experience, the motif encoding features are trained by large datasets, which has more usability, guidance quality and information. In the pooling layer, the output of the convolutional layer is reduced. And the most representative feature pattern of all motifs is selected as the feature representation. Finally the logistic regression function achieves the siRNA silencing efficacy prediction. Because too many layers produce a large number of weights, according to the scale of existing siRNA dataset, our CNN model only contains one convolutional layer and one pooling layer to detect the potential feature pattern of siRNA multimode motif, and one output layer to achieve siRNA silencing efficacy prediction.

B. The encoding method applicable to our CNN model

A siRNA sequence is composed of 21 bases, which can be expressed as:

$$S=s_1 \cdots s_i \cdots s_{21},\ s_i \in \{A,U/T,G,C\} \tag{1}$$

In this paper, to perform convolution operation, siRNA sequence was encoded into a two-dimensional matrix as the input of our CNN model. The four bases *A*, *U/T*, *G* and *C* were represented in four dimensional binary form, such as *A*=<1,0,0,0>, *U/T*=<0,1,0,0>, *G*=<0,0,1,0> and *C*=<0,0,0,1>. Then the siRNA sequence was transformed into 21×4 matrix. For example, S=*CUAAUAUGUUAAUUGAUUUat*, can be expressed as:

$S =$

$$\begin{bmatrix} 0&0&1&1&0&1&0&0&0&0&1&1&0&0&0&1&0&0&0&1&0 \\ 0&1&0&0&1&0&1&0&1&1&0&0&1&1&0&0&1&1&1&0&1 \\ 0&0&0&0&0&0&0&1&0&0&0&0&0&0&1&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix}^T$$

C. The motif detector design

As convolution operation can detect the local feature, we designed various convolution kernels to explore the potential feature pattern contained in siRNA multimode motifs. In the training stage of our CNN model, the weights of convolution kernels are corrected by large-scale training samples using back-propagation algorithm, which will ensure effective feature pattern are extracted.

According to the siRNA sequence encoding method, the siRNA sequence is transformed into 21×4 matrix and each base is represented by a four dimensional binary code. Because the size of convolution kernel cannot be greater than the size of input matrix, we used the convolution kernel of size $m \times 4$ as the motif detector. m represents the length of motif detector, and its value is range from 2 to 20. Various sizes of convolution kernels are used to detect the contributions of multimode motifs with different length to siRNA silencing efficacy prediction. The designed convolution kernel can detect the local features of multimode motif and the global features of whole sequence.

The convolution operation between $m \times 4$ convolution kernel M and siRNA sequence S is carried out and the neuron x_k of convolutional layer can be computed as followed:

$$x_k = \sum_{j=1}^{m} \sum_{i=1}^{4} \delta_k S_{k+j-1,i} M_{j.i} \tag{2}$$

$1 \le k \le 22-m$, δ_k is the learning rate of weight correcting. The size of convolutional result is $(22\text{-m}) \times 1$, which represents the feature pattern of each multimode motif.

Then ReLU function was used as the activation function to increase the nonlinear factors in our convolution layer. The comparison experiment showed that ReLU function performed best. The output y_k is shown as followed:

$$y_k = \max(0, x_k) \tag{3}$$

The pooling layer is used to choose the most representative convolution result and get rid of the irrelevant information. To keep the most prominent local feature

representation and the whole information of the convolution result, this paper implemented two choices for the pooling stage: max pooling and average pooling.

After pooling operation, the convolution results were reduced to a 2-dimensional vector $y = (y_{\max}, y_{avg})$.

$$y_{\max} = \max(y_1, \cdots, y_k) \tag{4}$$

$$y_{avg} = avg(y_1, \cdots, y_k) \tag{5}$$

Because different convolution kernels are needed to detect the contributions of the multimode motif with different size to siRNA silence efficiency, our convolution neural network output a 2d-dimensional vector for regression prediction, where d is the number of convolution kernel.

D. The logistic regression for siRNA silencing efficiency prediction

The activation function was implemented in the pooling layer. Since the quantitative siRNA silence efficiency prediction is a regression problem, there is only one neuron in the output layer. And we developed a logic regression model to realize the prediction process. The logic regression model is performing a linear weighting between the neurons of pooling layer and connecting weights wi, then mapping to the output value through the sigmoid function.

The input values in the range of $[-\infty, +\infty]$ are mapped into the values in the range of [0,1] by sigmoid function, which is consistent with the range of siRNA silencing efficacies. And the central area of sigmoid function has higher gain and both sides have lower gain, which is suitable to eliminate the errors caused by singular points in the feature vectors. The out layer is calculated as followed:

$$efficacy = sigmoid(\sum_{i=1}^{n} w_i h_i) \tag{6}$$

h_i is the output of pooling layer, and w_i is the connection weight.

E. The training process of siRNA silence efficiency prediction model based on our CNN algorithm

The gradient of convolution neural network is calculated by back-propagation algorithm. In this section, we estimated the error value according to the approximate minimization principle of loss function in output layer. Then the convolutional kernel of convolutional layer and the weight of output layer are iterative corrected.

The error between current output value and label is Δf. Because the activation function of output layer is sigmoid function, the gradient vector Δw of connected weight w_i in the pooling layer is shown as followed:

$$\Delta w_j = \Delta h_j h_j (1 - h_j) w \quad (1 \le j \le n) \tag{7}$$

$$\Delta w_{n+1} = \Delta f \tag{8}$$

We implemented max pooling and average pooling for the pooling stage. Since there are no model parameters, the feedback quantity is calculated as followed:

$$\Delta y_k = \begin{cases} \Delta w_{2k+1} & if\ i = argmax(y_1, \cdots, y_k) \\ 0 & otherwise \end{cases} + \frac{\Delta w_{2k}}{22 - m} \tag{9}$$

The convolutional layer is the first layer of forward propagation and the last layer of back propagation. The weights of convolution kernel are corrected according to the feedback quantity from the pooling layer. Since the activation function of convolutional layer is RELU function, the correction of neuron x_k of convolutional layer is calculated as followed:

$$\Delta x_k = \begin{cases} \Delta y_k & if\ y_k > 0 \\ 0 & otherwise \end{cases} \tag{10}$$

Thus the correction of weight of convolutional kernel is calculated as followed:

$$\Delta M_{k,l} = \sum_{i=1}^{22-m} S_{i+k-1,l} \Delta x_i \tag{11}$$

$1 \leqslant k \leqslant 22\text{-}m$, $1 \leqslant l \leqslant 4$.

F. Assessment of the prediction system

In this paper, Pearson Correlation Coefficient (PCC) and Receiver Operating Characteristic (ROC) curve are used to evaluate the performance of prediction model.

PCC describes the correlation between the observed and predicted siRNA activities of prediction model, and it can be defined as:

$$PCC = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right) \tag{12}$$

where n is the sample size. X_i and $\overline{X}$ are the observation value and average value respectively.

ROC curve is widely applied to compare the efficiencies of different algorithms in bioinformatics and generated by plotting sensitivity versus 1-specificity. Their definitions are as followed:

$$\text{Sensitivity} = TP/(TP+FN) \tag{13}$$

$$\text{Specificity} = TN/(TN+FP) \tag{14}$$

where TN is the number of true negatives; FN is the number of false negatives; TP is the number of true positives and FP is the number of false positives. The area under the

ROC curve (AUC) can be used to measure the overall performance of prediction algorithm.

III. RESULTS

A. Dataset

The dataset for comparative experiment includes 4067 siRNA samples, which are collected from Huesken dataset[11], Reynolds dataset[18], Vickers dataset[19], Haborth dataset[20], Takayuki dataset[21], Ui-Tei dataset[22] and siRNAdb dataset[23]. 10-fold cross-validation was used to select the optimal hyper parameter. In every experiment, 407 siRNA samples are randomly extracted as testing set and the remaining 3660 siRNA samples are as training set to develop the training model.

B. The hyper parameter setting

The hyper parameters of our convolutional network include the size of convolution kernel, activation function, learning rate and so on. These hyper parameters determine the network structure and directly impact the robustness of model. Since there is no prior knowledge to guide the setting of hyper parameters, the comparative experiments for all hyper experiments were designed to observe the impact of these hyper parameters on the prediction results and sum up the rules of the hyper parameter setting.

C. The impact of the size of convolution kernel on prediction result

In this section, we chose different sizes of convolution kernels to learn the feature representation of multimode motif with different lengths. Because the length of siRNA sequence is 21nt, there are 19 multimode motif with length from 2 to 20. Consequently, we employed the 19 convolutional kernels to learn the feature of multimode motif. The size of convolution kernel is $m\times 4$, and $2\leqslant m\leqslant 20$. In our experiment, 19 convolution neural networks are constructed according to the value of m, and there are 21-m+1 convolution kernels as the motif detector to learn the feature pattern of multimode motif in the corresponding convolution neural network. The experimental result is shown in fig 2.

As shown in fig.2, the size of convolution kernel influenced the prediction results. As the increasing of m, PCC and AUC predicted by the corresponding convolution neural network become larger. When m=15, PCC and AUC reached the highest value. The result showed that the effective feature pattern was learned from siRNA sequence by our proposed motif detector. The smaller size of the convolution kernel can only detect the information related to low-mode motif and ignore the contribution of high-mode motif and the global sequence feature. The large convolution kernel can cover more information of multimode motif and improve the prediction result. But if the convolution kernel is too large, the convolutional neural network will pay more attention to detecting the feature representation of high-mode motif and ignore the contribution of single nucleotide and low-mode motif with local significance, which bring about prediction result decrease. Therefore, reasonable convolution kernel size directly affects the effective motif feature learning. The convolution kernel which achieved PCC of 0.6 will be as the motif detector and ensure the learning feature pattern has enough discriminating ability.

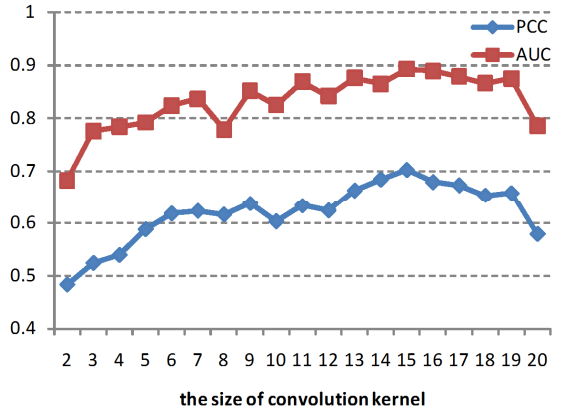

Figure 2. The impact of the size of convolution kernel on prediction result

D. The impact of the size of convolution kernel on prediction result

In this section, we compared the effect of different activation functions on the prediction result. Since the activation function affect the rationality and distinguish ability of the input signal mapping to the feature space, this section will examine the prediction effect of different activation functions. There are two activation functions in convolution layer and pooling layer. The common activation functions of convolution layer include tanh function, sigmoid function and RELU function. Because tanh function is more suitable to the situation that there is a big difference between feature values, it is not conducive to detecting the motif local tiny features. In this paper, sigmoid function and RELU function are suitable for the convolution layer, and sigmoid function and RELU function are suitable for the output layer. Fig.3 showed the prediction result between different activation functions in the convolution layer and polling layer.

Fig. 3 showed that the choice of the activation function has a considerable effect on the prediction effect. The best combination is that the activation function of convolution layer is ReLU function, and the activation function of output layer is sigmoid function. ReLU added the sparsity into the feature representation extracted from the convolution kernel, which could improve the amount of information contained in the nonzero neurons. And the pooling layer removed the neurons with zero values contained in the convolution result, which can guarantee the motif detector to get the most discriminative feature pattern of multimode motif. Consequently, ReLU is more suitable than sigmoid for the convolution layer. In the output layer, sigmoid can sum up the contribution of every feature component and output the values range from 0 to 1. And the output value of ReLU is

range from 0 to $+\infty$, which is not suitable as the active function of the output layer.

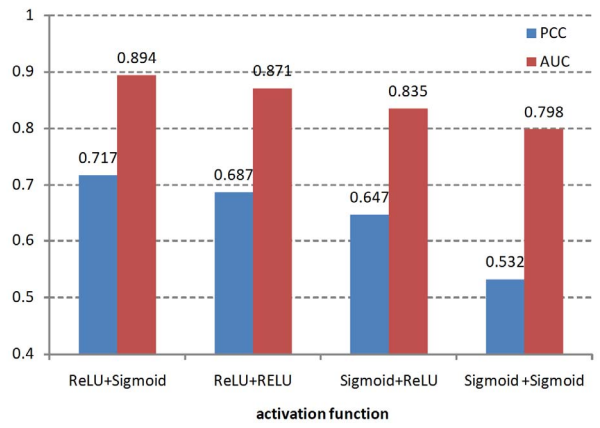

Figure 3. The impact of the activation function on prediction result

E. *The impact of learning rate on prediction result*

Learning rate is the coefficient controlling the corrected weight rate. The value of learning rate affects if the network weight can converge to the optimal network parameters. For different input signals, the adaptation learning rate needs to be adjusted experimentally. If the learning rate is too large, the network will miss the best network parameters and fall into local extremum. However, too little learning rate will cause slow convergence speed and insensitivity to error correction. In this experiment, we selected different learning rates for comparative experiments. If the iteration time is over 1000 or error is less than 0.001, the training can be terminated. According to our collected literature about the convolutional neural network, we selected 0.5, 0.1, 0.01 and 0.001 as the candidate set to analyze the influence of learning rate on prediction results. The experimental result is shown in Figure 4.

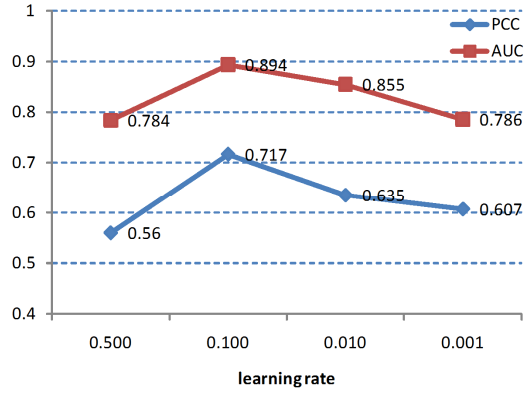

Figure 4. The impact of learning rate on prediction result

Fig.4 showed that when the learning rate is 0.1, the prediction result is best. And the prediction result is lowest when the learning rate is 0.5, indicating that the network training has missed the best weight and been plunged into the local extremum. When the learning rate is 0.01 and 0.001, the value of PCC and AUC are relatively low. This result showed that when the iterative is 1000, the network has slower convergence speed and not obtain the optimal weights. Considering the training time and prediction results, the learning rate of convolution neural network model is set to 0.1.

F. *Compared with other algorithms*

According to the above results, there are 14 convolution kernels with the size of 6×4 to 19×4 in our CNN model to detect the motif feature pattern. The activate function of convolutional layer is ReLU, the activation function of output layer is sigmoid, and the learning rate is 0.1. In this paper, we compared our model with other machine learning methods, including siRNApred[24], Biopredsi[11] and DSIR[14]. The PCC and AUC values of the five methods of were shown in Figure 5.

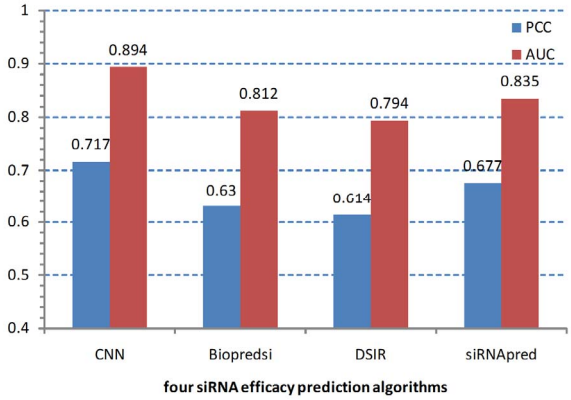

Figure 5. The comparison among four algorithms

Fig. 5 showed that the PCC of our method is 0.717, higher than Biopredsi, DSIR and siRNApred 13.81%, 16.78% and 5.91%. The result indicated the CNN model can effectively learn the potential feature pattern of multimode motif. Because the mapping weights are totally obtained by training data, it can more fully reflect the siRNA sequence feature and get a better prediction result.In addition, the AUC of our method is 0.894, which is higher than Biopredsi, DSIR and siRNApred 10.10%, 12.59% and 7.07%. The result shows that our method has a greater advantage for siRNA silencing efficiency prediction and is more stable and effective than other machine learning methods.

IV. CONCLUSIONS

siRNA has become a widespread molecular tool to study gene function and new identify drug target. Many efforts has been made to design active siRNAs. Therefore, it is important to develop an effective method to predict siRNA silencing efficacy. In this study, we developed a new siRNA efficacy prediction method based on Convolutional Neural Network. Comparing with the existing method Biopredsi, i-

Score, ThermoComposition21 and DSIR, our method shows better ability. The results showed that our model could explore the contribution of siRNA multimode motif on efficacy prediction and fully extract the feature pattern with the valuable information contained in sequence local feature. In light of this, the data-driven feature learning pattern has better performance than the learning pattern depending on the expert knowledge.

ACKNOWLEDGMENT

This research is partially supported by National Natural Science Foundation of China (11372155, 61402098) and the Scientific and Technological Development Program of Jilin Province (20170520058JH).

REFERENCES

[1] L. Timmons and A. Fire, "Specific interference by ingested dsRNA," *Nature,* vol. 395, p. 854, 1998.

[2] M. K. Montgomery, S. Xu, and A. Fire, "RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 95, pp. 15502-15507, 1998.

[3] S. M. Elbashir, "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells," *Nature,* vol. 411, pp. 494-498, 2001.

[4] C. D. Novina and P. A. Sharp, "The RNAi revolution," *Nature,* vol. 430, pp. 161-4, 2004.

[5] D. C. Baulcombe, "RNA as a target and an initiator of post-transcriptional gene silencing in trangenic plants," *Plant Molecular Biology,* vol. 32, p. 79, 1996.

[6] C. Cogoni, J. T. Irelan, M. Schumacher, T. J. Schmidhauser, E. U. Selker, and G. Macino, "Transgene silencing of the al-1 gene in vegetative cells of Neurospora is mediated by a cytoplasmic effector and does not depend on DNA-DNA interactions or DNA methylation," *Embo Journal,* vol. 15, pp. 3153-63, 1996.

[7] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans," *Nature,* vol. 391, p. 806, 1998.

[8] M. Elhefnawi, N. Hassan, M. Kamar, R. Siam, A. L. Remoli, I. El-Azab*, et al.*, "The design of optimal therapeutic small interfering RNA molecules targeting diverse strains of influenza A virus," *Bioinformatics,* vol. 27, pp. 3364-70, 2011.

[9] P. A. Sharp, "siRNA-directed inhibition of HIV-1 infection," *Nature Medicine,* vol. 8, pp. 681-686, 2002.

[10] P. Resnier, T. Montier, V. Mathieu, J. P. Benoit, and C. Passirani, "A review of the current status of siRNA nanomedicines in the treatment of cancer," *Biomaterials,* vol. 34, p. 6429, 2013.

[11] D. Huesken, J. Lange, C. Mickanin, J. Weiler, F. Asselbergs, J. Warner*, et al.*, "Design of a genome-wide siRNA library using an artificial neural network," *Nature Biotechnology,* vol. 23, pp. 995-1001, 2005.

[12] F. He, Y. Han, J. Gong, J. Song, H. Wang, and Y. Li, "Predicting siRNA efficacy based on multiple selective siRNA representations and their combination at score level," *Scientific Reports,* vol. 7, p. 44836, 2017.

[13] S. A. Shabalina, A. N. Spiridonov, and A. Y. Ogurtsov, "Computational models with thermodynamic and composition features improve siRNA design," *Bmc Bioinformatics,* vol. 7, Feb 12 2006.

[14] J. P. Vert, N. Foveau, C. Lajaunie, and Y. Vandenbrouck, "An accurate and interpretable model for siRNA efficacy prediction," *BMC Bioinformatics,* vol. 7, p. 520, 2006.

[15] M. Ichihara, Y. Murakumo, A. Masuda, T. Matsuura, N. Asai, M. Jijiwa*, et al.*, "Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities," *Nucleic Acids Research,* vol. 35, p. e123, 2006.

[16] F. He, Y. Han, H. Wang, J. Ji, Y. Liu, and Z. Ma, "Deep learning architecture for iris recognition based on optimal Gabor filters and deep belief network," *Journal of Electronic Imaging,* vol. 26, p. 023005, 2017.

[17] L. O. Chua and T. Roska, "CNN paradigm," *IEEE Transactions on Circuits & Systems I Fundamental Theory & Applications,* vol. 40, pp. 147-156, 1993.

[18] A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W. S. Marshall, and A. Khvorova, "Rational siRNA design for RNA interference," *Nature Biotechnology,* vol. 22, pp. 326-330, 2004.

[19] T. A. Vickers, S. Koo, C. F. Bennett, S. T. Crooke, N. M. Dean, and B. F. Baker, "Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis," *Journal of Biological Chemistry,* vol. 278, p. 7108, 2003.

[20] J. Harborth, S. M. Elbashir, K. Vandenburgh, H. Manninga, S. A. Scaringe, K. Weber*, et al.*, "Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing," *Antisense & Nucleic Acid Drug Development,* vol. 13, pp. 83-105, 2003.

[21] T. Katoh and T. Suzuki, "Specific residues at every third position of siRNA shape its efficient RNAi activity," *Nucleic Acids Research,* vol. 35, p. e27, 2007.

[22] K. Uitei, Y. Naito, F. Takahashi, T. Haraguchi, H. Ohkihamazaki, A. Juni*, et al.*, "Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference," *Nucleic Acids Research,* vol. 32, pp. 936-948, 2004.

[23] A. M. Chalk, R. E. Warfinge, P. Georgiihemming, and E. L. Sonnhammer, "siRNAdb: a database of siRNA sequences," *Nucleic Acids Research,* vol. 33, pp. D131-4, 2005.

[24] Y. Han, Y. Liu, H. Zhang, F. He, C. Shu, and L. Dong, "Utilizing Selected Di- and Trinucleotides of siRNA to Predict RNAi Activity," *Computational and Mathematical Methods in Medicine,2017,(2017-01-24),* vol. 2017, p. 5043984, 2017.