Front. Biol. China 2008, 3(2): 154–159
DOI 10.1007/s11515-008-0032-z

RESEARCH ARTICLE

Predicting siRNA activity based on back-propagation neural network

Jianlong LI, Zhengzhi WANG (✉), Xiaomin WANG

Institute of Mechanical Engineering and Automation, National University of Defense Technology, Changsha 410073, China

Abstract RNA interference (RNAi) is a phenomenon of gene silence induced by a double-stranded RNA (dsRNA) homologous to a target gene. RNAi can be used to identify the function of genes or to knock down the targeted genes. In RNAi technology, 19 bp double-stranded short interfering RNAs (siRNA) with characteristic 3′ overhangs are usually used. The effects of siRNAs are quite varied due to the different choices in the sites of target mRNA. Moreover, there are many factors influencing siRNA activity and these factors are usually nonlinear. To find the motif features and the effect on siRNA activity, we carried out a feature extraction on some published experimental data and used these features to train a back-propagation neural network (BP NN). Then, we used the trained BP NN to predict siRNA activity.

Keywords RNA interference (RNAi), double-stranded RNA (dsRNA), back-propagation neural network (BP NN)

1 Introduction

RNA interference (RNAi) is the degradation of homologous RNA induced by double-stranded RNA (dsRNA) in eukaryotic organisms (Denli and Hannom, 2003). In cells, long dsRNAs are cut into small interfering RNAs (or short interfering RNAs, siRNA) of 21–26 nucleotides (nt) by a nuclease in the RNase III family known as a Dicer. These siRNAs are incorporated into a protein complex to form an RNA-induced silencing complex (RISC) and then double-stranded siRNAs is unwound. The unwound anti-sense strands combine with the complementary transcripts (the target mRNA) which cause the degradation of the target mRNA. This is a post-transcriptional gene silence (PTGS), also known as RNA silence. The RNAi technique is a field with a broad application potential which may play an important role in the following aspects: (1) Applications in functional studies of genes. Currently, for a known gene, we can design the dsRNAs that can induce the silencing of this gene and then inject these dsRNAs into cells or organisms so as to reduce the gene expression level or to silence the gene completely. In this way, we can know the function of this gene. This is the widest application of the RNAi technique. Meanwhile, using the easily handled RNAi technique, we can silence large quantity of genes by batches in the genome, thus, making the examination of the function of a lot of genes in a shorter time possible; (2) Applications in disease therapy research. With the RNAi technique, we can specifically silence the endogenous or exogenous genes that are related to disease, or we can silence the protein genes related to the synthesis and decomposition of the intermediary metabolism in the process of the disease, while not interfering with the growth and development of the individual. This characteristic brings greater promise in the applications in disease therapy for RNAi; (3) Applications in plant antivirus research. Under natural conditions, when the abnormal replication of the virus in the plant is interrupted, dsRNA would be reproduced, which would cause the RNA interference in the plant and the degradation of the virus genome. Therefore, the RNA interference system is a natural defense system against virus. Using RNA interference, we can give the plants an antiviral ability. The principle is as follows, by injecting the dsRNAs homologous to the virus in the plant artificially, we can specifically cut and degrade the virus genome based on the RNA interference mechanism in the plant so as to prevent the replication and pervasion of the virus and to protect the plant from damage from the virus.

Translated from *Acta Biophysica Sinica*, 2006, 22(6): 429–434 [译自: 生物物理学报]

E-mail: wangzhengzhi@126.com

2 Materials and methods

2.1 siRNA activity

siRNA activity refers to its ability to silence target mRNA. The higher siRNA activity is, the more powerful

its ability to inhibit the expression of target mRNA. siRNA activity may be measured by the percentage of the reduction of target mRNAs or their protein products (compared with the status of no interference). Usually if the amount of target mRNAs or their protein products remains less than 50% after being interfered with, the siRNA is considered functional. Otherwise, it is considered non-functional.

2.2 The factors that influence siRNA activity

The RNAi technique requires that siRNA antisense strands have strict base pairing with the target gene sequences. The mispairing of a single base would enormously reduce the effect of silencing. It shows that different target sites choosing affect siRNA activity greatly; at certain sites the effect of siRNA silencing the target genes is very good, while at other sites the effect is poor and even does not work at all. Therefore, the design of siRNA sequences is of primary importance in the RNAi technique.

In order to design highly active siRNA sequences, Tuschl (Elbashir et al., 2001b) summarized the following basic principles in their study: (1) AA nucleotide sequences may be found downstream from the initiation codon AUG of the target gene transcripts. These AA nucleotide sequences and the downstream neighboring 19 nucleotides is the design template of siRNA sequences; (2) Avoid the region of poly-G ($\geqslant 4$) so as to prevent the inhibition of RNAi caused by tetramer structure; (3) G-C content of the chosen template sequence should be from 45% to 55%; (4) Avoid choosing target site within 50 to 100 nt from the initiation codon downstream or from the stop codon upstream of the mRNA and try not to use the sequences in the non-coding region of 5′ or 3′ end of mRNA as the design template of siRNA, because there are a lot of regulatory protein–binding sites (such as translation initiation complex) in these region and regulatory proteins would usually compete with RISV to combine with target sequences, there is, consequently, a reduction in the effect of gene silencing of siRNA; (5) For each gene, we can choose 4 to 5 siRNA sequences and then carry out homology analysis using bioinformatics and select the most specific siRNA while eliminating the sequences that are homologous to other genes.

Khvorova et al. (2003) discovered that the 5′-terminal antisense strand of active siRNA sequences was usually more unstable than that of comparatively inactive siRNA sequences and siRNA sequences with different activity had different internal free energy.

Schwarz et al. (2003) found that for siRNA sequences with high activity, the 5′-terminal antisense strand was more unstable than 3′-terminal.

RNA is a single strand nucleic acid consisting of four bases: A, C, G and U. The characteristics of single strand for RNA makes it different from DNA as part of the nucleotides in RNA can pair with other part of nucleotides in itself, i.e. RNA molecules can be folded into secondary or even a three-dimensional structure. Research shows that the secondary structure of mRNA could influence the RNA interference to a great extent (Vickers et al., 2003; Kawasakil et al., 2003). Among all the secondary structures, the multi-branch ring structure has a great influence on the accessibility of the target site on mRNA, that is, a complicated multi-branch ring structure may prevent siRNA sequences from combining to the target site on the mRNA.

On the other hand, different secondary structures of target mRNA have diverse stabilities (free energy) and its thermodynamic characteristic is quite different from that under the ideal status of a single strand. Therefore, the siRNA activity would be influenced. Considering the secondary structures of target mRNA, the main related parameters about free energy are: ΔG°_{37} (BTREAK-TARGET), ΔG°_{37} (OLIGO-SELF), ΔG°_{37} (OVERALL).

ΔG°_{37} (BTREAK-TARGET) refers to the free energy of the paired nucleotides in the corresponding areas of target mRNA. The smaller this value is (or the greater the absolute value is), the more stable the paired nucleotides in the corresponding areas of target mRNA are, and accordingly, the less effect siRNA can have.

ΔG°_{37} (OLIGO-SELF) refers to the structural free energy of antisense strand. If, after the unwinding of siRNA, the antisense strand can not form stable secondary structure, ΔG°_{37} (OLIGO-SELF) = 0. The more stable the formed secondary structure is, the greater the absolute value of ΔG°_{37} (OLIGO-SELF) would be, which means that it would be very difficult for siRNA to combine with the corresponding areas of target mRNA, i.e. siRNA has very low activity.

ΔG°_{37} is defined as follows:

$\Delta G^{\circ}_{37} = \Delta H^{\circ} - T\Delta S^{\circ} = \Delta H^{\circ} - 310.15\Delta S^{\circ}$, where 310.15 is in Kelvin degree which equals 37 Celsius degree (Freier et al., 1986).

Free energy is computed by the nearest neighbor model (NN) (Xia et al., 1998) and the parameters are shown in Table 1.

Table 1 Free energy values used

first nucleotide base pair	second nucleotide/kcal·mol^{-1}			
	A	C	G	U
A-U	−1.1	−2.4	−1.9	−1.1
C-G	−2.2	−3.3	−2.2	−1.9
G-C	−2.7	−3.8	−3.3	−2.4
G-U	−1.5	−2.7	−2.2	−1.4
U-A	−1.4	−2.6	−2.2	−1.1

2.3 The factors that influence siRNA activity

Artificial neural network (ANN) is a kind of machine learning method, which has great processing capacity toward complicated non-linear problems.

BP (back-propagation) algorithm, a kind of artificial neural network algorithm, is a training algorithm for non-cycling multistage network. Its basic principle is to estimate the error of the direct previous layer by the error of the output layer, and then to use this error to estimate the error of the layer prior to the previous layer. In this way, errors of all the layers could be estimated. Then we can adjust the weight of each node according to the errors and carry out recurrence until the precision of the output meets the requirement or the cycle index meets the predefined upper limit (Fig. 1) (Jiang, 2001).

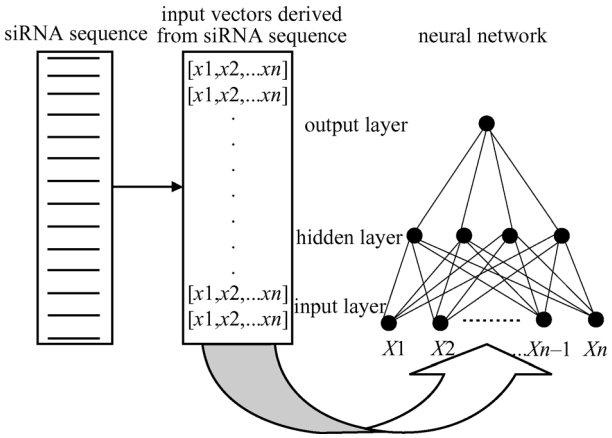

Fig. 1 Schematic diagram of the method for training and prediction

2.4 siRNA data preparation

All siRNA sequences and the activity data used in this study come from the published experimental data. We used 176 siRNA sequences of five different mRNA (GenBank J03132, U92436, M16553, M 60857, and U47298) (Khvorova et al., 2003; Vickers et al., 2003; Holen et al., 2002). The data of siRNA sequences were divided into two datasets_training set and testing set. The training set contains 146 siRNA sequences and the testing set contains 30 siRNA sequences. In order to avoid the influence of the sequence similarity between training set and testing set, the target sites distance between any two sequences from two sets must be no less than 50 bp.

2.5 Extracting eigenvectors

Now, we have already analyzed the possible factors that might influence siRNA activity. Since the free energy of the combination of A and T is greater than that of G and C, i.e., G-C is more stable than A-T. Therefore, we could use the counts of A and T in the first four nucleotides at 5′- and 3′-terminal siRNA antisense strand to show the thermodynamic characteristics of the two terminal siRNAs and use the counts of G and C between the eighth and twelfth nucleotide of siRNA antisense strand to show the stability of the mid position of siRNA sequence.

The mechanism of RNA interference indicates that after the unwound antisense strand combined with target mRNA, target mRNA would be cut into short fragments of 21 to 23 nt and the cut points are located near to the center of the area covered by siRNA. Therefore, the stability of base pairs at the center of siRNA sequences has the biggest influence on the cutting effect. Thus, we chose the feature vector *X5* to show this characteristic.

siRNA start to unwind from the 5′-terminal of antisense strand and then the unwound antisense strand combines with the target mRNA and finally causes the degradation of target mRNA. If the first base at the 5′-end of antisense strand is A-T, then it would be easier for siRNA to unwind. Thus, we chose the feature vector *X10* to show this characteristic.

The analysis shows that if there were ATG sequences in the siRNA, it was quite possible that this siRNA sequence had high activity, that is, the existence of ATG sequences was positively correlated with siRNA activity.

Feature vector *X11* shows the influence of RNA secondary structure on siRNA activity. When the target site of siRNA is located between two branches in a multi-branch ring (the number of branches is more than 3), *X11* equals L/N, where N refers to the number of nucleotides in the ring body of the multi-branches ring, and L refers to the number of nucleotides in the ring body between the two branches where the target site of siRNA is located. Intuitively, *X11* shows the extent of "crowdedness" of the target region. The smaller this value is, the tighter this target region is cracked by the two branches. Thus, the accessibility of this region would be very poor and siRNA would have little effect (Li and Wang, 2006). When the target region is not located in a multibranch ring, *X11* is set to 1.

Figure 2 shows a ring with 6 branches. N which refers to the number of nucleotides in the ring body is 27. The wave line in the figure denotes the target region and there are two nucleotides (denoted by the white circle in the figure) in the ring body between the two branches where this region is located, thus L equals 2.

After the extraction of eigenvectors of siRNA sequences in the training set and the testing set, there are altogether 11 vectors in the eigenvector set which reflect some factors that influence siRNA activity, as shown in Table 2.

2.6 Constructing and training BP neural network

Through many experiments and verifications, we finally decided to adopt the architecture shown in Fig. 1. The architecture of BP neural network was 10-8-6-1. We used the BP neural network tools provided by the software DPS (Data Processing System) to create the network with

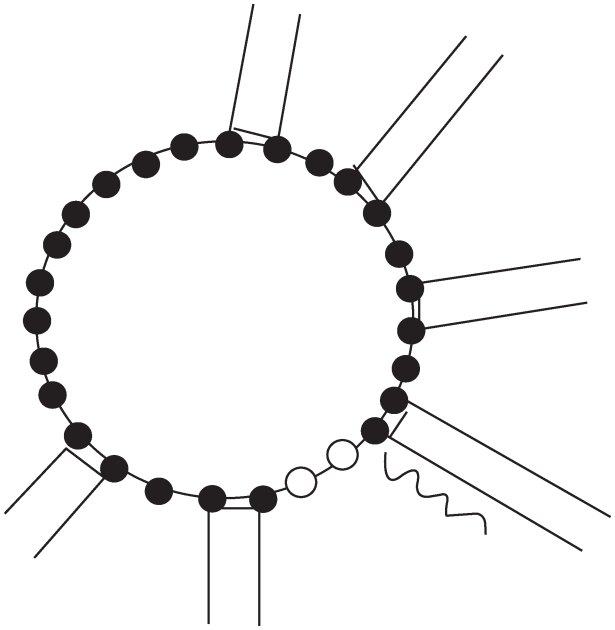

Fig. 2 A ring with 6 branches

the following parameters: the minimum training rate 0.1, SIGMOID parameter 0.9, dynamic parameter 0.6, permissible error 0.001 and the maximum number of iteration 2500. The process of error convergence in the course of training is shown as Fig. 3.

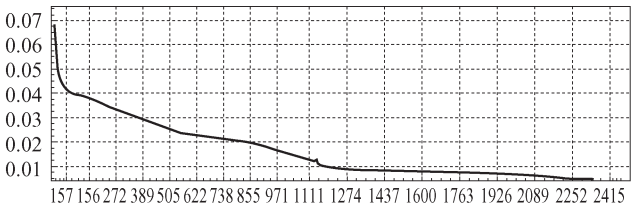

Fig. 3 Error profile of the ANN network training run. Sum squared error = 0.0043526775529823

3 Results

3.1 Evaluation criterion

Taking the inhibition rate of target mRNAs or their protein products caused by siRNA as the standard, siRNA sequences with an inhibition rate less than 50% are considered as nonfunctional, while those with an inhibition rate greater than 50% were considered as functional sequences. Obviously, the goal of prediction is to make sure that for a specified target mRNA, the rate of truly functional siRNA sequences in the predicted functional ones should be as high as possible and there should be a certain amount of siRNA sequences for the choosing.

Usually 4 indexes are used for evaluation: sensitivity (*Sn*), specificity (*Sp*), Matthews's coefficient (*M*), and average accuracy (*AC*).

Tp, *Tn*, *Fp* and *Fn* are defined as the number of truly functional, the number of truly nonfunctional siRNAs, the number of actually nonfunctional siRNAs and the number of actually functional siRNAs, respectively, which included in the predicted nonfunctional sequences. Then we can have the following equations:

$$Sn = \frac{Tp}{Tp + Fn}$$

$$Sp = \frac{Tp}{Tp + Fp}$$

$$AC = \frac{Tp + Tn}{Tp + Fp + Tn + Fn}$$

$$M = \frac{Tp.Tn - Fp.Fn}{\sqrt{(Tp + Fn)(Tn + Fp)(Tp + Fp)(Tn + Fn)}}$$

3.2 Prediction results

There are 30 siRNA sequences in the testing set among which 12 are functional. The other 18 are nonfunctional. The prediction result is that 9 sequences are functional,

Table 2 The parameter input sets

vector	signification
X1	GC content
X2	If contains ATG ?(yes: X2 = 1;no: X2 = 0)
X3	The quantity of AT in the first 4 nucleotides in the 5′ antisense region
X4	The quantity of AT in the first 4 nucleotides in the 5′ antisense region minus that in the first 4 nucleotides in the 3′ antisense region
X5	The 10th nucleotide of antisense region (if AT:X5 = 0;if GC:X5 = 1)
X6	The quantity of GC in the antisense region from the 8th to the 12th nucleotide
X7	If contains GGGG or CCCC?(yes: X7 = 1;no: X7 = 0)
X8	ΔG°_{37} (OLIGO-SELF)
X9	ΔG°_{37} (BTREAK-TARGET)
X10	The first nucleotide of the 5′ antisense region (if AT:X10 = 0;if GC:X10 = 1)
X11	L/N

among which 8 are really functional. 21 sequences are nonfunctional, among which 17 are really nonfunctional. Therefore:

$$Sp = \frac{Tp}{Tp+Fp} = \frac{8}{8+4} = 0.67$$

$$Sp = \frac{Tp}{Tp+Fp} = \frac{8}{8+1} = 0.89$$

$$AC = \frac{Tp+Tn}{Tp+Fp+Tn+Fn} = \frac{8+17}{30} = 0.83$$

$$M = \frac{Tp.Tn - Fp.Fn}{\sqrt{(Tp+Fn)(Tn+Fp)(Tp+Fp)(Tn+Fn)}} = \frac{8*17-1*4}{\sqrt{(8+4)*(17+1)*(8+1)*(17+4)}} = 0.65$$

From the prediction result, we can see that the probability of predicted functional siRNA sequences being really functional nearly reaches 89%. Meanwhile, the parameter *Sn* is 0.67, which shows that on one hand, this method can guarantee high reliability of the predicted functional siRNA sequences. On the other hand, it can make sure that a certain proportion of the truly functional candidate siRNA sequences could be predicted so as to guarantee that there are a certain amount of functional siRNA sequences for the choosing in experiments or applications.

4 Discussion

Regarding the problem on how to design effective siRNA sequences, the above mentioned Tuschl etc. first brought forward several basic principles, known as Max-Planck-Institute (MPI) criterion which has been extended and developed by some one else. But this criterion is still experimental.

In 2004, Pål first used the machine learning method to predict siRNA activity. He examined siRNA activity using Genetic Programming (GP) and Support Vector Machine (SVM) and the algorithms derived from the two methods, also gave the experimental results. The dataset of siRNA sequences adopted in this paper is similar to that of Pål, thus our method is comparable with their method to some extent, and the comparison result is shown in Table 3 (Pål did not give parameter *AC*, so this parameter is not compared; the parameter *Se* in Pål (2004) is the same to *Sn* in this paper).

As shown in Table 3, our method is better than the other two methods in each index, which is mainly due to the more feature extractions we have done for siRNA sequences. Essentially, we can catch more factors that influence siRNA activity. As a result, our method can improve the accuracy of recognition remarkably.

Table 3 Prediction results based on different algorithm

algorithm	*Sn*	*Sp*	*M*
BP ANN (our work)	0.67	0.89	0.65
GP	0.45	0.73	0.19
GPboost	0.56	0.70	0.27
GPboostReg	0.50	0.73	0.24
v-SVM (seq)	0.52	0.56	0.09
C-SVM (seq)	0.35	0.70	0.05
v-SVM (1–2)	0.61	0.68	0.31
C-SVM (1–2)	0.45	0.80	0.26
v-SVM (4)	0.49	0.73	0.25
C-SVM (4)	0.35	0.82	0.19

5 Conclusions

This method utilized a machine learning method called BP neural network to integrate all kinds of factors that influence siRNA activity, such as the sequence characteristics, thermodynamic characteristics and secondary structure characteristics of the siRNA sequence. The experimental results show that our method can be effectively applied in the design of siRNA sequence.

Meanwhile, this method has a good extensibility. With the deepening of the understanding toward the mechanism of RNAi, the newly discovered factors that influence siRNA activity could be added into the input vector of the BP neural network conveniently so as to further improve the accuracy of the recognition.

Acknowledgements The work was supported by the National Natural Science Foundation of China (Grant No. 60471003).

References

Denli A M, Hannom G J (2003). RNAi: an ever-growing puzzle. Trends Biochem Sci, 28: 196–201

Elbashir S M, Lendeckel W, Tuschl T (2001b). RNA interference is mediated by 21-and 22-nucleotide RNAs. Genes Dev, 15: 188–200

Freier S M, Kierzek R, Jaeger J A, Sugimoto N, Caruthers M H, Neilson T, Turner D H (1986). Improved free-energy parameters for prediction of RNA duplex stability. Proc Natl Acad Sci USA, 83: 9373–9377

Holen T, Amarzguioui M, Wiiger M T, Babaie E, Prydz H (2002). Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. Nucleic Acid Res, 30: 1757–1766

Jiang Z L (2001). Introduction to artificial neural network. Beijing: Higher Education Press (in Chinese)

Kawasakil H, Suyama E, Iyo M, Taira K (2003). siRNAs generated by recombinant human Dicer induce specific and significant but target site-independent gene silencing in human cells. Nucleic Acid Res, 31: 981–987

Khvorova A, Reynolds A, Jayasenal S D (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell, 115: 209–216

Li J L, Wang Z Z (2006). Researches on the relationship of siRNA activity and mRNA. Biomedical Engineering, 25: 46–49

Pål S (2004). Predicting the efficacy of short oligonucleotidesin antisense and RNAi experiments with boosted genetic programming. Bioinformatics, 20: 3055–3063

Schwarz D S, Hutvagner G, Du T, Xu Z, Aronin N, Zamore P D (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell, 115: 199–208

Vickers T A, Koo S, Bennet C F, Crooke S T, Dean N M, Baker B F (2003). Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. J Biol Chem, 278: 7108–7118

Xia T, Santa Lucia J J, Burkard M E, Kierzek R, Schroeder S J, Jiao X, Cox C, Turner D H (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry, 37: 14719–14735