Gene 562 (2015) 152–158

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Soft computing model for optimized siRNA design by identifying off target possibilities using artificial neural network model

Reena Murali [a,*], Philips George John [a], David Peter S [b]

[a] Department of Computer Science and Engineering, Rajiv Gandhi Institute of Technology, Kerala, India
[b] Department of Computer Science and Engineering, Cochin University of Science & Technology, Kerala, India

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form 25 January 2015
Accepted 22 February 2015
Available online 25 February 2015

Keywords:
Small interfering RNA
Artificial neural network
Messenger RNA
Area Under Curve
Off target

ABSTRACT

The ability of small interfering RNA (siRNA) to do posttranscriptional gene regulation by knocking down targeted genes is an important research topic in functional genomics, biomedical research and in cancer therapeutics. Many tools had been developed to design exogenous siRNA with high experimental inhibition. Even though considerable amount of work has been done in designing exogenous siRNA, design of effective siRNA sequences is still a challenging work because the target mRNAs must be selected such that their corresponding siRNAs are likely to be efficient against that target and unlikely to accidentally silence other transcripts due to sequence similarity. In some cases, siRNAs may tolerate mismatches with the target mRNA, but knockdown of genes other than the intended target could make serious consequences. Hence to design siRNAs, two important concepts must be considered: the ability in knocking down target genes and the off target possibility on any non-target genes. So before doing gene silencing by siRNAs, it is essential to analyze their off target effects in addition to their inhibition efficacy against a particular target. Only a few methods have been developed by considering both efficacy and off target possibility of siRNA against a gene. In this paper we present a new design of neural network model with whole stacking energy (ΔG) that enables to identify the efficacy and off target effect of siRNAs against target genes. The tool lists all siRNAs against a particular target with their inhibition efficacy and number of matches or sequence similarity with other genes in the database. We could achieve an excellent performance of Pearson Correlation Coefficient ($R = 0.74$) and Area Under Curve ($AUC = 0.906$) when the threshold of whole stacking energy is ≥ -34.6 kcal/mol. To the best of the author's knowledge, this is one of the best score while considering the "combined efficacy and off target possibility" of siRNA for silencing a gene. The proposed model shall be useful for designing exogenous siRNA for therapeutic applications and gene silencing techniques in the area of bioinformatics. The software is developed as a desktop application and available at http://opsid.in/opsid/

1. Introduction

In functional genomic research, the discovery of RNA interference (RNAi) has become much helpful in drug design and therapeutic applications because of its ability to perform gene silencing. The RNAi pathway was discovered by Fire and Mello in 1998 (Fire et al., 1998). RNAi is a biological process which can control the gene regulation by a sequence specific posttranscriptional gene silencing mechanism (Elbashir et al., 2001) mediated by double stranded RNA (dsRNA). RNAi has been successfully used to target diseases such as AIDS (Martínez Miguel et al., 2002), neurodegenerative diseases (Xia et al., 2004), cholesterol (Soutschek et al., 2004) and cancer (Borkhardt, 2002) on mice with the hope of extending these approaches to treat humans. The RNAi can be endogenous or exogenous. The use of exogenous siRNA for performing gene silencing has become an important biological milestone for drug design and mRNA target identification (McManus and Sharp, 2002; Hannon and Rossi, 2004). A significant amount of work has been done over the recent past to understand the gene silencing mediated by siRNA. Hence designing siRNA against target mRNA or gene (Hannon and Rossi, 2004; Meister and Tuschl, 2004), with good knockdown efficiency and target specificity is an area of concern to be addressed upon.

Even though several algorithms and methods have been developed to predict efficiency of siRNA, only a few of them have achieved an acceptable level of specificity and sensitivity. These algorithms are classified into two groups, first generation and second generation methods. As the first generation tools were not able to achieve the targeted level of efficacy, there was a need to develop techniques to improve the efficiency of predicted siRNA. These second generation

Abbreviations: siRNA, small interfering RNA; miRNA, microRNA; mRNA, messenger RNA; RNAi, RNA interference; ROC, receiver operating characteristics; AUC, Area Under Curve; dsRNA, double stranded RNA; ANN, artificial neural network; OpsiD, Optimized siRNA Designer; MCC, Matthews Correlation Coefficient

* Corresponding author.
E-mail addresses: reena.rajesh@rit.ac.in, reena.raajesh@gmail.com (R. Murali).

http://dx.doi.org/10.1016/j.gene.2015.02.067
0378-1119/

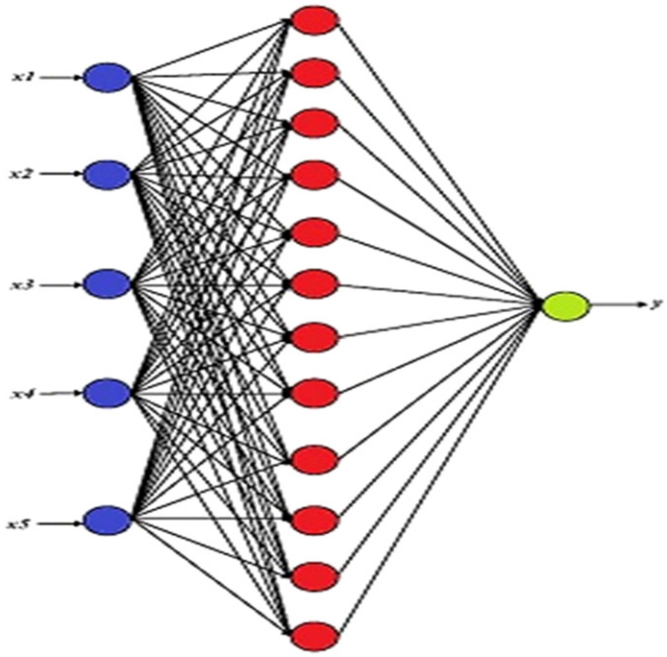

Fig. 1. 5-12-1 neural network model.

models are based on either artificial neural network or linear regression models. Some of the good scoring second generation tools like Biopredsi (Huesken et al., 2005), DSIR (Vert et al., 2006), ThermoComposition21 (Shabalina et al., 2006), i-Score (Ichihara et al., 2007), Scales (Matveeva et al., 2007), My siRNA-Designer (Mysara et al., 2011) and MysiRNA (Mysara et al., 2012) were developed by introducing data mining techniques to improve the efficiency of siRNA with their experimental inhibition. Biopredsi, ThermoComposition21, MysiRNA-Designer package and MysiRNA used the artificial neural network models, while DSIR, i-Score and Scales used linear regression models. ThermoComposition21 improved the prediction accuracy by combined position dependent features together with thermodynamic features in single artificial neural network model. The prediction accuracy is improved in DSIR, i-Score and Scales using linear regression model. Further the MysiRNA-Designer package and MysiRNA much improved the prediction accuracy by artificial neural network model.

But from the studies related to siRNA, it is understood that among all siRNAs that can be generated against a target mRNA, only a fraction are successful in causing degradation and all siRNAs do not perform equal knockdown effects (Holen et al., 2002). The efficiency of siRNA differs in different target sites of the same mRNA. Thus the aim of siRNA efficiency prediction is to design siRNA sequences that are highly capable of inhibiting their target mRNA sequences. Earlier it was understood that full complementary siRNA is needed to silence a target gene. But recent studies reveal that siRNA behaves like miRNA and siRNA can suppress protein synthesis even though it is not fully complementary to the target. This shows that mismatches are allowed during target selection by siRNA (Ameres et al., 2007; Doench et al., 2003). This may cause a very serious problem of off-target effect where unintended target genes may be suppressed by selected siRNA (Burchard et al., 2009; Jackson et al., 2003, 2006). In this work, we propose an Optimized siRNA Designer and Off target Finder (OpsiD) which use some of the known functionalities and thermodynamic features like whole stacking energy (ΔG) as parameters and produce efficient siRNAs with high inhibition efficacy to degrade target genes or mRNAs, and identify the off target knockdown possibilities of siRNA against unintended genes.

Table 2
Sample siRNAs with inhibition capacity and BLAST score in our tool.

siRNA strand	Inhibition	BLAST score
AGGGUUAUUUUUCUUUGGC	75	11
GAAAAAAACCAAAGGGUUA	67	3
AACCACUGUAGAAAAUAAC	35	0
UCUUUAUGUUUUUGGCGUC	89	17
UUCUUUAUGUUUUUGGCGU	76	9
GGGCCUUUCUUUAUGUUUU	55	7
UUAUAAAUGUCGUUCGCGG	77	12
UAAUUUUUUGGAUGAUUGG	45	4
UUAAAAUCGCAGUAUCCGG	67	8

2. Materials and methods

2.1. Neural network

A multi-layer perceptron feed-forward neural network, trained using the Scaled Conjugate Gradient training algorithm provided by Encog (Moller, 1993) is used for computing the final score of each siRNA. The Scaled Conjugate Gradient algorithm is based upon a class of optimization techniques well known in numerical analysis as the Conjugate Gradient Methods. The neural network consists of an input layer with 5 neurons, a single hidden layer with 12 neurons and output layer with 1 neuron which can be viewed as a 5-12-1 neural network and is shown in Fig. 1.

For selecting the neural network model, we tested various configurations of feed-forward neural networks such as 4-8-8-1, 5-7-7-1, 5-8-1, 5-8-8-1, 5-10-1, 5-12-1, 6-7-7-1, 6-8-8-1, 6-8-8-8-1, 6-10-1 and 6-12-1. A number of neural network training algorithms such as the classic Backpropagation, Resilient propagation (RProp) and Scaled Conjugate Gradient (SCG) were tried out. Varying numbers of training iterations were also tried out depending on the network configuration and the training algorithm. We finally chose a configuration of 5-12-1 which, according to our observations, gave the best balanced result in terms of Pearson's correlation coefficient, sensitivity, specificity and ROC with the training and testing data sets, and which did not overfit the training data set. The training was done for 1,36,000 iterations, and started from a randomized state. The neural network was built and trained using the Encog Workbench IDE and later integrated into our siRNA designer tool OpsiD.

2.2. Input values

For computing the final score of each siRNA, we considered five different metrics: whole stacking energy (ΔG), DSIR score (Vert et al., 2006), ThermoComposition21 score (Shabalina et al., 2006), i-Score prediction value (Ichihara et al., 2007), and MysiRNA score (Mysara et al., 2012). In our experiments, these metrics were found to work well and gave good results with the Huesken data set (Huesken et al.,

Table 1
Data Sets used for training and testing our model.

Train/test Data Set	Name of Data Set	No of genes	No of siRNA used	Source of siRNA	siRNA with 50% to 70% inhibition	siRNA with 70% to 90% inhibition	siRNA with >90% inhibition
Train Data Set	Data Set 1	30	2431	Huesken et al. (2005)	778	853	369
Test Data Set	Data Set 2	12	419	Ichihara et al. (2007)	60	117	96
Test Data Set	Data Set 3	9	476	Mysara et al. (2012)	70	53	127