Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 336 (2005) 723–728
BBRC
www.elsevier.com/locate/ybbrc

Prediction of siRNA knockdown efficiency using artificial neural network models

Guangtao Ge [a,b,*], G.William Wong [b], Biao Luo [c]

[a] Department of Computer Science, Tufts University, 161 College Avenue, Medford, MA 02155, USA
[b] Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA
[c] RNAi Consortium, Broad Institute, Massachusetts Institute of Technology, 320 Bent Street, Cambridge, MA 02142, USA

Received 2 August 2005
Available online 29 August 2005

Abstract

Selective knockdown of gene expression by short interference RNAs (siRNAs) has allowed rapid validation of gene functions and made possible a high throughput, genome scale approach to interrogate gene function. However, randomly designed siRNAs display different knockdown efficiencies of target genes. Hence, various prediction algorithms based on siRNA functionality have recently been constructed to increase the likelihood of selecting effective siRNAs, thereby reducing the experimental cost. Toward this end, we have trained three Back-propagation and Bayesian neural network models, previously not used in this context, to predict the knockdown efficiencies of 180 experimentally verified siRNAs on their corresponding target genes. Using our input coding based primarily on RNA structure thermodynamic parameters and cross-validation method, we showed that our neural network models outperformed most other methods and are comparable to the best predicting algorithm thus far published. Furthermore, our neural network models correctly classified 74% of all siRNAs into different efficiency categories; with a correlation coefficient of 0.43 and receiver operating characteristic curve score of 0.78, thus highlighting the potential utility of this method to complement other existing siRNA classification and prediction schemes.

Keywords: siRNA; Knockdown efficiency; Artificial neural network

RNAi (RNA interference) represents an extremely powerful approach to silence gene expression using synthetic, plasmid or viral-encoded small interfering RNAs (siRNAs) [1]. Guided by RNA induced silencing complex (RISC), siRNA binds to its complementary target mRNA and induces its degradation [2]. Since its discovery [3], RNAi technology has been widely used to study and validate gene function through selective knockdown of their target mRNAs. Thus, gene function can be inferred by comparing the phenotypes or functional differences before and after the introduction of siRNA specific for a given gene of interest. The simplicity and ease with which this technology can be applied has made it a powerful tool to interrogate mammalian genome in a high throughput manner to uncover novel pathways involved in diseases [4].

siRNA typically consists of a short RNA sequence with 19 nucleotides and a 3′ 2-nt T overhang. Potential siRNAs are selected from each target mRNA by sliding a 19 nucleotide-long window along its entire length. Based on certain siRNA functional features, several siRNAs will then be chosen, synthesized, and empirically tested for their ability to knockdown target gene expression. Not surprisingly, the efficacy of each siRNA spans the range of no effect to near complete knockdown of target gene expression. Thus, any method that increases the chances of selecting an effective siRNA will greatly reduce the experimental cost and validation time involved. Toward this end, several groups have recently developed some general rules, based on experimental data, to select good candidate siRNA. In a systematic analysis of 180 siRNAs targeting the mRNA of two genes,

* Corresponding author. Fax: +1 617 627 3220.
E-mail addresses: guge@eecs.tufts.edu (G. Ge), wong@wi.mit.edu (G.William Wong), bluo@broad.mit.edu (B. Luo).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.08.147

Reynolds et al. [5] discovered eight functional features associated with an effective siRNA. These include low GC content, a bias towards low internal stability at the 3′-terminal on the sense strand, and other base preferences. In a separate study, Amarzguioui et al. [6] analyzed the ability of 46 siRNAs to knockdown the expression of four genes and found six features that correlated well with siRNA functionality. Both groups agreed that the asymmetry in the stability of the duplex end correlated well with siRNA functionality, but disagreed on the contribution of specific sequence motif. Other studies also employed similar approaches to select the best candidate siRNA for a given gene [7–9]. In general, one point is assigned to each functional feature, and each siRNA is then scored according to how many functional features it possesses. Any siRNA with a score above a user-defined threshold will then be selected. This scoring procedure makes two important assumptions: each feature of the siRNA is independent of each other, and all the features are equally important in its overall ability to silence gene expression.

Attempts were also recently made to apply various machine learning algorithms to help select effective siRNA candidates with high knockdown efficiencies of target genes from a much larger test set of siRNA database. Saetrom [10] implemented a boosted genetic programming algorithm on an assembled database of 204 siRNAs. Based on their best model using sequence pattern encoding method, they observed an overall correlation coefficient of 0.46 between the predicted and observed siRNAs knockdown values of target gene expression, with an receiver operating characteristic (ROC) score of 0.72. Of the published reports, this prediction method has the highest correlation value. Using only highly correlated features of siRNA functionality based on 287 siRNAs from 30 genes, Chalk et al. [11] could achieved more than a twofold improvement in prediction over random selection of siRNA on a trained regression tree with full cross-validation. Recently, support vector machines based on simplified generalized string kernel and other kernel methods also have been used in this context [12,13].

Neural network models are a form of machine learning technique that can effectively handle noise and complex relationship in a more robust way. Among the more extensively studied neural network models include the back-propagation neural network (BPNN), general regression neural network (GRNN), and probabilistic neural network (PNN). The prediction capabilities of these neural network models have been proved successfully in other contexts [14–17]. Here, we demonstrated the utility of using three different neural network models, comparable to the best published method, to predict and classify siRNA into different knockdown efficiency categories.

Materials and methods

Neural network models. Artificial neural network is built on a set of interconnected neural units and consists of one input and one output layer that takes the input values and outputs the final output result individually.

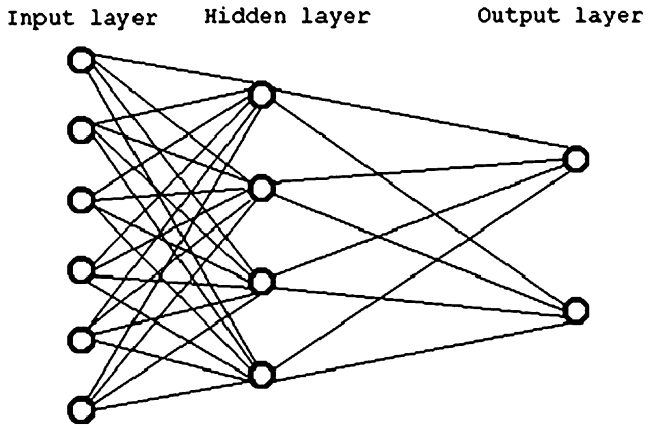

Fig. 1. Back-propagation neural network model with full connections between different layer units. The graph shows there is one hidden layer in between the input and output layer.

Some of them have one or more hidden layers which perform nonlinear modeling (Fig. 1).

There are many different types of neural networks. Each differs from the others in network topology and/or learning algorithm. In this study, we introduce the back-propagation, general regression, and probabilistic neural networks in the context of predicting siRNA knockdown efficiency.

Back-propagation neural network. Back-propagation neural network is a multilayer feed-forward network with hidden layers between the input and output layer. Each unit in hidden layer calculates a weighted net output of the input units.

$$\mathrm{net}_j = \sum_{i=1}^{n} x_i w_{ij} + w_0 = \sum_{i=0}^{n} x_i W_{ij}, \tag{1}$$

where net_j is the value of jth unit in hidden layer and x_i is the value of ith unit in input layer. W_{ij} is the weight of connection between these two units. W_0 is the bias. Through certain activation functions, the output unit produces class labels.

$$f(\mathrm{net}) = \mathrm{Sgn}(\mathrm{net}) = \begin{cases} 1 & \text{if net} \geqslant 0, \\ -1 & \text{if net} < 0, \end{cases} \tag{2}$$

where Sgn is signal function and here we assume 0 is the threshold we used. Used in supervised learning task, the network assigns random weights to all interconnections between units initially and computes the error term between desired target value and prediction. It then propagates the error backward through the network. Thus by approximating the desired target value, the network iteratively adjusts these weights to find the best combination of weights to minimize the prediction error.

Specific learning algorithms have been previously developed for back-propagation neural network. The learning procedure iteratively presents all patterns to the classifier and adjusts those weights until the error term is below or equal to the user-defined threshold. By adjusting free parameters such as learning time, the number of hidden units, learning rate, and momentum, the algorithm searches the error space and tries to find the local minimum, which yields the best classification result.

General regression neural network. General regression neural network (GRNN) [18] is another form of feed-forward network. It differs from back-propagation neural network in several ways. First, all GRNNs have four layers: input, hidden, summation, and output layers. The number of units in hidden layer and summation layer depend on the number of patterns and input vector. Second, GRNNs use radial basis function instead of sigmoid activation function to describe the target probabilistic value. It estimates the joint probability density function using Parzen's nonparametric window. Third, GRNNs use a one-pass learning process. Therefore, it is faster compared to back-propagation algorithm. Expectation of the target value given an input is described by Eq. (3):