Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6476–6487
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Using radial basis function networks and significance testing to select effective siRNA sequences

Shigeru Takasaki*, Yoshihoro Kawamura

RIKEN Genomic Sciences Center (GSC), Suehiro-cho 1-7-22-E216, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan

Received 30 June 2006; received in revised form 5 January 2007; accepted 26 February 2007
Available online 12 March 2007

Abstract

Although short interfering RNA (siRNA) has been widely used for studying gene functions in mammalian cells, its gene silencing efficacy varies markedly and there are only a few consistencies among the recently reported design rules/guidelines for selecting siRNA sequences effective for mammalian genes. We propose a method for selecting effective siRNA target sequences by using a radial basis function (RBF) network and statistical significance analysis for a large number of known effective and ineffective siRNAs. The siRNA classification is first carried out by using the RBF network and then the preferred and unpreferred nucleotides for effective siRNAs at individual positions are chosen by significance testing. The gene degradation measure is defined as a score based on the preferred and unpreferred nucleotides. The effectiveness for the proposed method was confirmed by evaluating effective and ineffective siRNAs for the recently reported genes (15 genes, 196 sequences) and comparing the scores thus obtained with those obtained using other scoring methods. Since the score is closely correlated with the degree of gene degradation, it can easily be used for selecting high-potential siRNA candidates. The evaluation results indicate that the proposed method may be applicable for many other genes. It will therefore be useful for selecting siRNA sequences in mammalian genes.

Keywords: siRNA design; RNA interference; Gene silencing; RBF network; Statistical significance

1. Introduction

Although RNA interference (RNAi) has been successfully used for studying gene functions in both plants and invertebrates, many practical obstacles need to be overcome before it becomes an established tool for use in mammalian systems (Fire et al., 1998; Sharp, 2001; Elbashir et al., 2001a, b; Hannon, 2002; Dykxhoorn et al., 2003). One of the important problems is designing effective siRNA sequences for target genes. The short interfering RNA (siRNA) responsible for RNA varies markedly in its gene silencing efficacy in mammalian genes, where the gene silencing effectiveness depends very much on the target sequence positions (sites) selected from the target gene (Elbashir et al., 2001c; Holen et al., 2002). Since different siRNAs synthesized for various positions induce different levels of gene silencing, the selection of the target sequence is critical to the effectiveness of the siRNA. We therefore need useful criteria for gene silencing efficacy when we design siRNA sequences (Mittal, 2004; Kumar et al., 2003).

Zamore et al. and Jayasena et al. showed that 5′ end of the antisense strand might be incorporated into RNA-induced silencing complex (RISC). Strand incorporation may depend on weaker base pairing and thus an A–T terminus may lead

* Corresponding author. Tel.: +81 45 503 9618; fax: +81 45 503 9559.
E-mail address: takasaki@gsc.riken.jp (S. Takasaki).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.02.027

to more efficient strand incorporation than a G–C terminus (Schwarz et al., 2003; Khvorova et al., 2003). Other factors reported to be related to gene silencing efficacy are GC content, point-specific nucleotides, specific motif sequences and secondary structures of mRNA. Several siRNA design rules/guidelines using efficacy-related factors have been reported (Chalk et al., 2004; Naito et al., 2004; Santoyo et al., 2004; Teramoto et al., 2005; Truss et al., 2005).

Although the positional nucleotide characteristics for siRNA designs seem to be the most important factor determining effective siRNA sequences, there are few consistencies among the reported rules/guidelines (Reynolds et al., 2004; Ui-Tei et al., 2004; Amarzguioui and Prydz, 2004; Hsieh et al., 2004; Jagla et al., 2005; Huesken et al., 2005). This implies that these rules/guidelines might result in the generation of many candidates and thus make it difficult to extract a few for synthesizing siRNAs. Furthermore, there is in RNAi a risk of off-target regulation: a possibility that the siRNA will silence other genes whose sequences are similar to that of the target gene. When we use gene silencing for studying gene functions, we have to first somehow select high-potential siRNA candidate sequences and then eliminate possible off-target ones (Snove et al., 2004).

Here we therefore focus on identifying high-potential siRNA sequences from many possible candidates and propose an effective method for selecting effective siRNA target sequences by using the radial basis function (RBF) technique and statistical significance analysis for a large number known effective and ineffective siRNAs. After the siRNA classification is carried out by using a RBF network (Poggio and Girosi, 1990; Wu and McLarty, 2000), the preferred and unpreferred nucleotides for effective siRNAs at individual positions are determined by significance testing and used to calculate a score that measures a sequence's potential for gene degradation. The effectiveness of the proposed scoring method was confirmed by using it to evaluate RNA sequences recently reported to effectively or ineffectively suppress the expression of various genes (see later sub-section) and comparing it with other scoring methods (Saetrom and Snove, 2004). We found a good correlation between the size (value) of the proposed score and the effectiveness and ineffectiveness of the recently reported siRNA sequences. The evaluation results indicate that the proposed method would be useful for many other genes. It will therefore be useful for selecting siRNA sequences for mammalian genes.

2. Systems and methods

An RBF network is a type of artificial network for application to problems of supervised learning, such as regression, classification and time series prediction. In the supervised learning, the function is learned from the examples (training set) which a teacher supplies. The training set contains elements which consist of paired values of the independent (input) variable and the dependent (output) variable. The RBF network is a new technique for value prediction that demonstrates more robustness and flexibility than traditional regression approaches such as neural networks and polynomial fits. The RBF network works by choosing not just a single nonlinear function, but a weighted sum of a set of nonlinear functions. These weighted functions are the so-called RBFs. The RBFs are each fitted to separate regions in the input space. The regions are chosen such that the output is quite similar within a region, so that the RBF is most likely to fit well to the output. For each selected region, an RBF center is created that predicts the average of the region. Data points that fall between regions are predicted by taking a weighted average of the predictions of all centers, where the weight for a center decays rapidly if the center is very far from that data point (Poggio and Girosi, 1990; Wu and McLarty, 2000). The RBF networks are supervised learning models with a single middle layer of units. They are similar to back propagation neural networks but usually faster to train because the RBFs used in the units mean that fewer weight adjustments are needed. Also, RBF networks tend to be more resistant to noisy data than back propagation networks.

2.1. Preparation

The relations between siRNA sequences and the effectiveness of their gene silencing are shown in Fig. 1A. For simplicity, sense strands of siRNAs (cDNA $5'$ to $3'$, 19 nucleotides from positions 1 to 19) are described. To use an RBF network for selecting effective siRNA sequences, we need to represent individual nucleotides (A, G, C and T) as numerical data. We therefore transform the symbols A, G, C and T into the following numerical representations: $A = 1$, $G = 2$, $C = 3$ and $T = 4$. Other numerical data representations for individual nucleotides are, of course, also possible. That is, if individual nucleotides were expressed by other numerical values, e.g., $A = 4$, $G = 3$, $C = 1$ and $T = 2$, the RBF network could accomplish similar classifications. The original set of siRNAs is thus transformed into a set of numerical values as shown in Fig. 1B. To make the RBF network for siRNA selections easily understood, we