Journal of
Computer Science & Systems Biology

Takasaki, J Comput Sci Syst Biol 2014, 7:1
DOI: 10.4172/jcsb.1000137

Research Article Open Access

Efficient Methods for Selecting Sirna Sequences by Using the Average Silencing Probability and a Hidden Markov Model

Shigeru Takasaki*

Toyo University, 1-1-1 Izumino Itakura-machi, Ora-gun Gunma, 374-0193, Japan

Abstract

Short interfering RNA (siRNA) has been widely used for studying gene functions in mammalian cells but varies markedly in its gene silencing efficacy. Although many design rules/guidelines for effective siRNAs based on various criteria have been reported recently, there are only a few consistencies among them. This makes it difficult to select effective siRNA sequences in mammalian genes. This paper first clarifies problems of the recently reported siRNA design guidelines and then proposes a new method for selecting effective siRNA sequences from many possible candidates by using the average silencing probability on the basis of large number of known effective siRNAs. It is different from the previous score-based siRNA design techniques and can predict the probability that a candidate siRNA sequence will be effective. The results of evaluating it by applying it to recently report effective and ineffective siRNA sequences for various genes indicate that it would be useful for many other genes. The evaluation results indicate that the proposed method would be useful for many other genes. It should therefore be useful for selecting siRNA sequences effective for mammalian genes. The paper also describes another method using a Hidden Markov Model (HMM) to select the optimal functional siRNAs.

Keywords: siRNA; siRNA design; RNA interference; Gene silencing; Estimation of gene silencing; Average gene silencing; Hidden Markov model

Introduction

RNA interference (RNAi) silences gene expression by introducing double-stranded RNA homologous to the target mRNA. It has been widely used for studying gene functions, but many practical obstacles need to be overcome before it becomes an established tool for use in mammalian systems [1-6]. One of the important problems is designing effective siRNA sequences for target genes. The effectiveness of the short interfering RNA (siRNA) responsible for RNA interference varies widely depending on the target sequence positions (sites) selected from the target gene [7,8]. We therefore need useful criteria for gene silencing efficacy when we design siRNA sequences [9,10].

Schwarz et al. and Khvorova et al. [11,12] showed that 5' end of the antisense strand might be incorporated into the RNA-induced silencing complex. Strand incorporation may depend on weaker base–pairing, and an A-T terminus may thus lead to more strand incorporation than a G-C terminus. 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 [13-17].

Although the effectiveness of siRNA sequences seems to be determined largely by their nucleotide sequences, there are few consistencies among the reported rules/guidelines [18-23]. This implies that they might result in the generation of many candidate sequences, making it difficult to select the effective ones. In addition, the previously reported rules/guidelines cannot estimate the probability that a candidate siRNA will actually silence the target gene. What are therefore needed are not only methods for selecting high-potential siRNA candidates but also methods for estimating the probability that the selected candidates will indeed silence their target genes. 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 [24].

This paper first reviews the recently reported siRNA design guidelines and clarifies their problems. It then describes a prediction method for selecting effective siRNA target sequence from many possible candidate sequences by using the average silencing probability of a large number of siRNA sequences known to be effective. It is quite different from the previous score-based siRNA design techniques and can predict the probability that a candidate siRNA sequence will be effective. The results obtained when applying the method to recently report effective and ineffective siRNA sequences for various genes showed that it is accurate and thus imply that it would be useful for selecting siRNA sequences silencing many other genes. Since siRNA sequences consisting of 19 nucleotides can be expressed as state diagrams of nucleotide A, C, G, or T from positions 1 to 19, they can be considered a hidden Markov process. If the state diagrams of effective siRNAs were expressed as HMM, the optimal states maximizing the transition probability could be solved by using the Viterbi algorithm. Therefore this paper describes another method using HMM to select the optimal functional siRNAs [25].

siRNA sequence selection problems

To use RNAi as a biological tool for mammalian cell experiments, we first need to identify target sequences causing gene degradation.

***Corresponding author:** Shigeru Takasaki, Toyo University, 1-1-1 Izumino Itakura-machi, Ora-gun Gunma 374-0193 Japan, Tel: +81-276-82-9024; Fax: +81-276-82-9033; E-mail: s_takasaki@toyo.jp

Received December 27, 2013; **Accepted** January 12, 2013; **Published** January 14, 2014

Citation: Takasaki S (2014) Efficient Methods for Selecting Sirna Sequences by Using the Average Silencing Probability and a Hidden Markov Model. J Comput Sci Syst Biol 7: 045-053. doi:10.4172/jcsb.1000137

They have so far been identified by using a trail-and-error method [3,8], but siRNAs extracted from different regions of the same gene have varied remarkably in their effectiveness. The difficulty of using the trail-and-error method to select target sequences causing gene silencing increases when the coding regions are long, as they are in mammalian cells. This is because the number of candidates increases with the length of the coding region.

The reported guidelines for designing siRNA sequences

The earliest guidelines for siRNA sequence design were proposed by Elbashir et al. [4,8,26]. They suggested that the target mRNA is silenced effectively by siRNA duplexes 21 nucleotides long: 19-nt base-paired sequences with 2-nt overhangs at the 3' ends. Many siRNA design guidelines/rules have been reported since then, and this paper treats the following five (here designated guidelines G1–G5).

Reynolds et al. [18] analyzed 180 siRNAs systematically, targeting every other position of two 197-base regions of firefly luciferase and human cyclophilin B mRNA (90 siRNAs per gene), and reported eight criteria for improving siRNA selection.

Guideline G1: (1) G/C content 30–52%), (2) at least 3 As or Ts at positions 15–19, (3) absence of internal repeats, (4) an A at position 19, (5) an A at position 3, (6) a T at position 10, (7) a base other than G or C at position 19, (8) a base other than G at position 13.

Ui-Tei et al. [19] examined 72 siRNAs targeting six genes and reported four rules for effective siRNA designs.

Guideline G2: (1) an A or T at position 19, (2) a G or C at position 1, (3) at least five T or A residues from positions 13 to 19, (4) no GC stretch more than 9 nt long.

Amarzguioui and Prydz [7] analyzed 46 siRNAs targeting four genes and reported six rules for effective siRNA designs.

Guideline G3: (1) a G or C at position 1, (2) an A at position 6, (3) a base other than T at position 10, (4) a T at position 13, (5) a C at position 16, (6) an A or T at position 19.

Jagla et al. [22] tested 601 siRNAs targeting one exogenous and three endogenous genes and reported four rules.

Guideline G4: (1) an A or T at position 19, (2) an A or T at position 10, (3) a G or C at position 1, (4) more than three A/Ts between positions 13 and 19.

Hsieh et al. [21] examined 138 siRNAs targeting 22 genes and reported five position-specific characteristics:

Guideline G5: (1) a T at position 19, (2) a C or G at position 11, (3) a G at position 16, (4) an A at position 13, (5) a base other than C at position 6.

These guidelines are summarized in Table 1.

Other methods for scoring, screening, and designing functional siRNAs have also been reported recently. Chalk et al. [13] reported the following seven rules ("Stockholm rules") based on thermodynamic properties: (1) total hairpin energy < 1, (2) antisense 5' end binding energy < 9, (3) sense 5' end binding energy in range 5-9 exclusive, (4) GC between 36% and 53%, (5) middle (7-12) binding energy < 13, (6) energy difference < 0, (7) energy difference between −1 and 0. The score of a siRNA candidate is incremented by one for each rule fulfilled and is thus between 0 and 7).

	Position	1	3	6	10	11	13	16	19
G1	effective		A		T		A/C/T		A/T
G2	effective	G/C							A/T
	ineffective	A/T							G/C
G3	effective	G/C		A			T	C	A/T
	ineffective	T			T				G
G4	effective	G/C			A/T				A/T
G5	effective					C/G	A	G	T
	ineffective			C		A/T			G

position: nucleotide position from 1 to 19 (5' to 3', cDNA form)
effective: preferred, ineffective: unpreferred

Table 1: Effective and ineffective nucleotides specified in the individual guidelines.

Huesken et al. [23] reported a method for screening functional siRNAs by using an artificial neural network. This network was first trained by 2182 randomly selected siRNAs targeted to 34 genes and was used in the design of a genome-wide siRNA collection with two potent siRNAs per gene.

Teramoto et al. [14] and Ladunga [29] have reported functional siRNA selection methods using Support Vector Machines (SVMs). Teramoto et al. [14] used a Generalized String Kernel (GSK) combined with a SVM. siRNA sequences were represented as vectors in a multidimensional feature space according to the numbers of subsequences in each siRNA and were classified as effective or ineffective. Ladunga used a SVM with polynomial kernels and constrained optimization models from 572 sequence, thermodynamic, accessibility, and self-hairpin features over 2200 published siRNAs [23,29]. As the key to SVM success is to collect many useful features of effective siRNA sequences, the usefulness of methods using SVMs may depend on the selected siRNAs.

Holen [27] recently reported siRNA rules based on apparent overrepresentation or underrepresentation of certain nucleotides in certain positions of Novartis data set. The criteria for a siRNA candidate depend on the positive and negative scores computed for each position by using scoring table generated by the percentage overrepresentation or underrepresentation of individual nucleotides for each position in the large Novartis data set [23]. Although the method was evaluated by using other reported siRNA sets, which of the candidate siRNAs actually silence genes is not clear. In addition, as the original scores in the scoring table are based on the percentage overrepresentation or underrepresentation of certain nucleotides in certain positions, they may vary drastically depending on what sets of siRNAs are used. This makes it difficult to evaluate the scores computed for siRNA candidates.

Although secondary structures of siRNA sequences are also thought to be important in predicting siRNA efficacy, there are conflicting results concerning the effects of secondary structures on siRNA functionality. Some studies have suggested that the secondary structure of the siRNA plays a role in determining the efficacy of gene silencing [33-35], but others did not find any correlation between the functionality of the siRNA and the secondary structures of the target mRNA [7,18,20]. This issue therefore requires further study.

The above techniques have also been used to obtain other design rules [36-40], and the features of various siRNA design rules are summarized in Table 2.

Problems with the previous guidelines

Among the problems with the reported guidelines is the problem of inconsistencies with regard to the nucleotide frequencies of

	No. of genes	No. of siRNAs	Description	Technique
Reynolds et al. [18]	2	197	Sequence features	
Ui-Tei et al.[19]	6	72	Sequence features	
Amarzguioui et al. [20]	4	46	Sequence features	
Hsieh et al. [21]	22	138	Sequence features	
Hesken et al. [23]	34	2128	Sequence motifs	Neural network
Jagla et al. [22]	4	601	Sequence features	Decision tree
Holen [27]	34	400	Sequence features	Percentage
Saetrom [28]	40	581	Sequence motifs	Genetic programming
Teramoto et al. [14]	2	94	Sequence motifs	Support vector Machine
Ladunga [29]	34	2252	Position features	Support vector Machine
Calk et al. [13]	92	398	Binding energy	Regression tree
Takasaki et al. [30-32]	490	833	Sequence features	Statistics, SOM, RBF

Table 2: Features of individual siRNA design rules.

each position. Although some guidelines have the same preferred and unpreferred nucleotides at positions 1 and 19, there are few consistencies at other positions (Table 1). These results indicate that though some rules from the guidelines are suitable for getting effective sequences for some genes, they might be unsuitable for others. Since the previous guidelines are based on the analyses of specific genes, it could be inferred that they are not always effective for many other genes. Therefore if these guidelines were used to select sequence candidates for other mammalian genes, many sequences might be selected as candidates. This is because there are mostly long coding regions in mammalian genes but there are only a few consistencies among the previous guidelines. As a result, many candidate sequences might be selected. This is another problem because experimentally evaluating whether the selected sequences provide effective gene degradation is a costly and time-consuming task. To overcome the problems of the previous guidelines, Takasaki et al. [41-43] recently reported new scoring methods using the statistical and clustering techniques listed in Table 2.

Still another problem is that the previously reported methods cannot estimate the probability that a candidate siRNA will actually silence the target gene. Even if a high-scored siRNA were obtained using the reported methods, it would be difficult to estimate the probability that it would actually accomplish the expected gene degradation.

Methods and Materials

Definition of the siRNA sequence selection problem

The problem of selecting target siRNA sequences is to predict whether or not a candidate siRNA sequence for the target mRNA (typically 19 nucleotides $\mathbf{X}=X_1, X_2, \ldots\ldots, X_{19}$ where X_i is the i-th nucleotide) will result in effective gene silencing: for example, more than 80% silencing at the protein level. The problem of selecting siRNA sequences can therefore be transformed into the problem of finding the degree of gene silencing functionality of a given siRNA candidate **X**. If individual gene reduction degrees of siRNA candidates were obtained, it might be easy to decide what candidates are appropriate as siRNAs.

It is hypothesized that the evaluation of candidates can be based on the analyses of nucleotide occurrence features at individual positions in the reported effective siRNAs. This is because the effectiveness of siRNA sequences greatly depends on individual nucleotides of the sequences [18-22].

Prediction analysis based on the average silencing probability

Many effective gene-silencing siRNA sequences have been reported recently and can be used to predict how new siRNA candidates will function. If the probability of individual nucleotide occurrences at positions from 1 to 19 in the effective siRNA population is obtained, it can be used to calculate the probability that candidate siRNAs will be effective. In addition, if the average probability of a large number of effective siRNA sequences is computed, it could be considered a measure of the potential effectiveness of siRNAs and used to evaluate whether or not a candidate siRNA is likely to silence its target gene. If the probability of the candidate siRNA candidate were greater than the average probability of a large number of effective siRNAs, it would indicate a high likelihood of gene-silencing. To calculate this measure, 833 effective siRNA sequences reported in the literature (PubMed) were collected and nucleotide occurrences at positions from 5' to 3' in the cDNA were summarized. The probability of individual nucleotide occurrence frequencies f_p^N at individual positions can be computed as follows:

$$f_p^N = \frac{\sum_{i=1}^{I} \{A, G, C, T\}}{I}, \quad (1)$$

where N is the kind of nucleotide (A, G, C, or T), p is the position in the cDNA (1, 2, …,19 from 5' to 3'), and I is the number of the effective siRNA sequences (e.g., 833).

Then the probability $\boldsymbol{OF_i}$ of each effective siRNA sequence is calculated in the following way.

$$OF_i = \prod_{p=1}^{19} f_{ip}^N, \quad (2)$$

where i is the sequence identification number of the effective siRNAs (i.e., i=1, 2, …, I).

The average sequence probability $\boldsymbol{A_E}$ for the effective siRNAs is therefore computed as follows:

$$A_E = \frac{\sum_{i=1}^{I} OF_i}{I} \quad (3)$$

$\boldsymbol{A_E}$ could be considered a criterion for candidate siRNA. That is, if the probability of the candidate sequence were greater than $\boldsymbol{A_E}$, it would indicate a high likelihood of gene-silencing. On the other hand, if the probability of the candidate sequence were remarkably lower than $\boldsymbol{A_E}$, it would indicate a low likelihood of effectiveness.

siRNA sequence selection based on a hidden Markov model

As a siRNA sequence $\boldsymbol{X}$ basically consists of 19 nucleotides, it can be described as $\mathbf{X}=X_1X_2\ldots\ldots X_{19}$, where X_i indicates the nucleotide A, C, G, or T. Furthermore, this sequence can be expressed as state diagrams of nucleotides A, C, G, and T from the positions 1 to 19 shown in Figure 1. As shown in Figure 1, if the state at position 1 is, for example, the nucleotide C, it can be transmitted to all the states A, C, G, or T at position 2. Likewise, these nucleotide state transitions proceed from the positions 1 to 19. In relations between the state diagrams (top) and the frequency ratios (bottom) as shown in Figure 1, although what states are allocated to the individual positions of effective siRNA sequences are unknown in the intermediate processes, the ratios of the individual nucleotide occurrences are obtained as shown in the bottom of Figure 1. Therefore, the transmission of the individual nucleotides A, C, G,

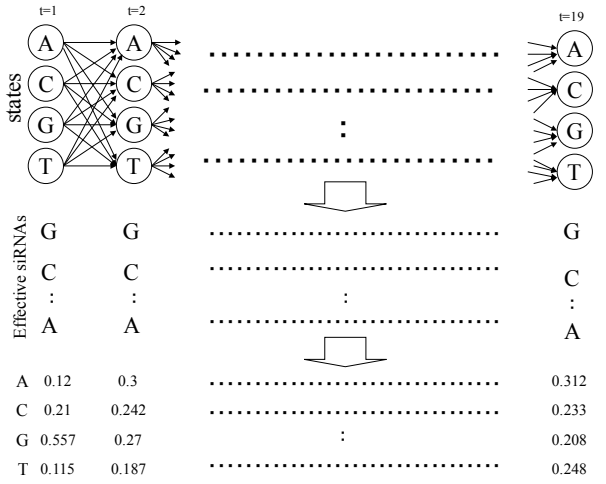

Figure 1: State diagram of hidden Markov model.A set of 833 effective siRNA sequences from the literature is shown in the middle part of the Figure 1, and the frequency ratios of individual nucleotides at each position in those sequences are shown in the bottom of Figure 1. The ratios of the nucleotides A, C, G, and T at position 2, for example, are respectively 250/833 (=0.3), 202/833 (=0.242), 225/833 (=0.27), and 156/833 (=0.187).

and T from the positions 1 to 19 can be considered a hidden Markov process. If the state diagrams of effective siRNAs were expressed as a hidden Markov model (HMM), the optimal states (nucleotides) for maximizing the state transition probability could be solved as a decoding problem by using the Viterbi algorithm (Figure 1).

Viterbi algorithm for selection of the optimal siRNA nucleotide

The Viterbi algorithm for selecting the optimal siRNA sequence is expressed as follows:

1. Initialization for individual states *i*=A, C, G, T.

$$\delta_1 = \prod_i b_i(O_1) \tag{4}$$

$$\varphi_1(\boldsymbol{i}) = 0$$

where $\prod_i$ is initial state probability distribution for the state *i* and $b_i(o_1)$ is output of the state *i* at the sequence position 1.

2. Recursive computations for the sequence positions *t* = 1, 2,…, 18 and the individual states *j* = A, C, G, T.

$$\delta_{t+1}(j) = \max_i \left(\delta_t(i)a_{ij}\right) b_j(O_{t+1}) \tag{5}$$

$$\varphi_{t+1}(j) = \arg\max_i \left(\delta_t(i)a_{ij}\right) \tag{6}$$

where a_{ij} is the state transition probability from the state *i* to *j*.

3. Termination of the recursive computations.

$$\hat{P} = \max_i \delta_{19}(i) \tag{7}$$

$$\hat{q}_{19} = \arg\max_i \delta_{19}(i) \tag{8}$$

4. Optimal state generation for sequence positions *t* = 18, 17, …, 1.

$$\hat{\boldsymbol{q}}_{\boldsymbol{t}} = \varphi_{t+1}(\hat{\boldsymbol{q}}_{t+1}) \tag{9}$$

Nucleotide occurrence models

Two types of nucleotide occurrences from positions 1 to 19 were assumed. One is that the nucleotides occur independently at individual positions as listed in Table 3a, and the other is that the occurrence of individual nucleotides at individual positions depends on the nucleotides at other positions. A typical occurrence dependency is, for example, the Markov chain dependency (the simple (first) Markov model). That is, the nucleotide occurrences at the present position depend on the nucleotides at the previous position. The probability of the simple Markov model for the nucleotide at the present position *i* (i=2, 3, …, 19) is determined under the condition of the effective nucleotide at the previous position *i-1* as listed in Table 3b. Suppose, for example, that we have the sequence CGACTGACGACGCAGATCT as a candidate siRNA sequence for some target gene. In this case, the probability of the second nucleotide being G may depend on the first nucleotide being C. From Table 3b (the simple Markov Model table) one sees that the probability that G occurs at position 2 under the condition that there is a 0.208 probability that C occurs at position 1 is 0.116. One similarly sees that the probability that the third nucleotide is A is 0.364 under the condition that there is a 0.27 probability that the second nucleotide is G (Table 3a-3c).

Evaluation criteria used in the proposed method

To make the results estimated using the average silencing probability easily understood, the ratio of the result estimated for a new siRNA candidate to the average sequence probability $\boldsymbol{A}_E$ is considered. This is because the results estimated for the known effective siRNAs could be considered a standard criterion for candidate new siRNAs. This normalized ratio ***NR*** is therefore defined as follows:

$$NR = \frac{ER}{A_E} \tag{10}$$

Where ***ER*** is the result estimated by the average silencing probability method and $\boldsymbol{A}_E$ is the average of the probabilities predicted for the known effective siRNAs.

This ***NR*** therefore indicates the gene-silencing potential of the siRNA candidates relative to that of the known effective siRNAs. If ***NR*** ≥ 1, the level of gene silencing expected to be obtained with the siRNA candidate is the same as or higher than level of silencing obtained with the known effective siRNAs. That is, ***NR*** indicates that the candidate sequence is likely to silence its target gene. If, on the other hand, ***NR***<1, the gene silencing expected to be obtained with the candidate sequence is lower than the level of silencing obtained with the known effective siRNAs.

Evaluation and model generation data

The recently reported effective and ineffective siRNAs were used as the evaluation data. They are respectively 25 effective and 25 ineffective sequences for human *cyclophilin B* [18]; 38 effective and 24 ineffective sequences for *firefly luciferase (PRL-TK), vimentin, Oct 4, EGFP, ECFP,* and *DsRed* [19]; 21 effective and 25 ineffective sequences for *hTF*, *mTF*, *PSK*, and *CSK* [20]; 7 effective and 7 ineffective sequences for the *cyclin B1* [42]; and 12 effective and 12 ineffective sequences for *TC10*, *UBE2I*, and *CDC34* [23]. These sets of genes are respectively symbolized throughout the present study as MG1, MG2, MG3, MG4, and MG5.

Two kinds of known effective siRNA sequences were used for obtaining frequency ratios of individual nucleotides. One was 833 effective siRNA sequences from 490 different cDNAs in the published

(a) Probabilities of independent nucleotide occurrences in 833 effective siRNAs

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	0.12	0.3	0.318	0.218	0.271	0.294	0.247	0.298	0.271	0.229	0.25	0.283	0.259	0.282	0.27	0.232	0.288	0.313	0.312
G	0.557	0.27	0.229	0.291	0.262	0.239	0.304	0.253	0.224	0.257	0.279	0.247	0.255	0.276	0.242	0.287	0.24	0.239	0.208
C	0.208	0.242	0.208	0.294	0.248	0.208	0.263	0.229	0.244	0.275	0.256	0.234	0.247	0.217	0.202	0.251	0.235	0.178	0.233
T	0.115	0.187	0.245	0.197	0.218	0.259	0.186	0.22	0.261	0.239	0.216	0.235	0.239	0.224	0.286	0.23	0.236	0.27	0.248

(b) Probabilities of dependent nucleotide occurrences – the simple Markov model

		1'-2		2'-3		3'-4		4'-5		5'-6		6'-7		7'-8		8'-9		9'-10	
A	A	0.12	0.16	0.3	0.328	0.318	0.2	0.218	0.247	0.271	0.292	0.294	0.229	0.247	0.282	0.298	0.278	0.271	0.204
	G		0.35		0.272		0.336		0.341		0.265		0.363		0.311		0.286		0.323
	C		0.28		0.192		0.264		0.22		0.212		0.257		0.214		0.254		0.257
	T		0.21		0.208		0.2		0.192		0.23		0.151		0.194		0.181		0.217
G	A	0.557	0.33	0.27	0.364	0.229	0.241	0.291	0.314	0.262	0.394	0.239	0.342	0.304	0.316	0.253	0.313	0.224	0.337
	G		0.284		0.169		0.293		0.194		0.206		0.241		0.241		0.218		0.193
	C		0.213		0.227		0.251		0.26		0.22		0.246		0.213		0.223		0.203
	T		0.172		0.24		0.215		0.231		0.179		0.171		0.229		0.246		0.267
C	A	0.208	0.358	0.242	0.371	0.208	0.277	0.294	0.331	0.248	0.329	0.208	0.295	0.263	0.37	0.229	0.33	0.244	0.261
	G		0.116		0.149		0.127		0.171		0.13		0.185		0.137		0.131		0.192
	C		0.295		0.198		0.318		0.245		0.203		0.243		0.224		0.199		0.271
	T		0.231		0.282		0.277		0.253		0.338		0.277		0.269		0.34		0.276
T	A	0.115	0.198	0.187	0.167	0.245	0.172	0.197	0.146	0.218	0.137	0.259	0.144	0.186	0.187	0.22	0.153	0.261	0.134
	G		0.396		0.353		0.368		0.409		0.368		0.389		0.361		0.246		0.304
	C		0.25		0.218		0.353		0.268		0.192		0.301		0.284		0.301		0.359
	T		0.156		0.263		0.108		0.177		0.302		0.167		0.168		0.301		

10'-11		11'-12		12'-13		13'-14		14'-15		15'-16		16'-17		17'-18		18'-19	
0.229	0.267	0.25	0.274	0.283	0.195	0.259	0.245	0.282	0.264	0.27	0.218	0.232	0.249	0.288	0.3	0.313	0.3
	0.356		0.284		0.314		0.329		0.306		0.329		0.275		0.288		0.235
	0.257		0.25		0.263		0.236		0.179		0.258		0.218		0.154		0.231
	0.12		0.192		0.229		0.19		0.251		0.196		0.259		0.254		0.235
0.257	0.271	0.279	0.358	0.247	0.364	0.255	0.373	0.276	0.33	0.242	0.267	0.287	0.285	0.24	0.37	0.239	0.387
	0.271		0.263		0.204		0.274		0.226		0.231		0.238		0.2		0.166
	0.285		0.19		0.204		0.193		0.165		0.249		0.247		0.185		0.196
	0.173		0.19		0.228		0.16		0.278		0.151		0.151		0.245		0.236
0.275	0.306	0.256	0.305	0.234	0.282	0.247	0.325	0.217	0.331	0.202	0.222	0.251	0.421	0.235	0.352	0.178	0.412
	0.131		0.155		0.169		0.146		0.16		0.107		0.148		0.133		0.108
	0.262		0.239		0.241		0.214		0.204		0.187		0.167		0.204		0.23
	0.301		0.3		0.308		0.316		0.304		0.231		0.263		0.306		0.23
0.239	0.146	0.216	0.172	0.235	0.204	0.239	0.181	0.224	0.144	0.286	0.151	0.23	0.188	0.236	0.228	0.27	0.183
	0.382		0.294		0.321		0.357		0.262		0.396		0.307		0.325		0.272
	0.213		0.267		0.281		0.226		0.273		0.236		0.313		0.173		0.263
	0.256		0.267		0.194		0.236		0.321		0.276		0.193		0.274		0.277

(c) Probabilities of independent nucleotide occurrences in 847 ineffective siRNAs

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	0.312	0.247	0.211	0.247	0.254	0.231	0.273	0.26	0.235	0.295	0.251	0.243	0.226	0.235	0.203	0.261	0.262	0.182	0.084
G	0.185	0.229	0.256	0.262	0.237	0.259	0.236	0.254	0.286	0.279	0.231	0.257	0.323	0.266	0.293	0.26	0.255	0.301	0.319
C	0.215	0.253	0.296	0.285	0.266	0.321	0.283	0.253	0.298	0.242	0.269	0.256	0.247	0.244	0.289	0.269	0.262	0.319	0.426
T	0.288	0.272	0.236	0.207	0.243	0.189	0.208	0.234	0.182	0.184	0.248	0.243	0.204	0.255	0.215	0.21	0.221	0.198	0.171

Table 3: Probabilities of individual nucleotide occurrences at each position.

references of the PubMed database [5,6,43] and the other was the 636 top-ranked effective siRNA sequences (normalized inhibitory activity >0.832) from 34 genes [23]. The reason is that various kinds of effective gene silencing siRNAs can be obtained from papers in PubMed. 833 effective siRNAs were used to calculate a standard criterion for the effectiveness of siRNAs. The evaluation data were not included in the 833 effective siRNAs. Since it is difficult to select many known ineffective siRNAs from many different genes, the 847 worst-ranked siRNAs (normalized inhibitory activity <0.612) from Huesken et al. [23] were used as ineffective siRNA sequences.

Results and Discussion

The proposed method was evaluated by first computing A_E for 833 effective siRNA sequences and then using equation (2) to compute the individual probabilities of the effective and ineffective siRNAs. Since there were ups and downs in the individual ratios of the effective and ineffective siRNAs, the average of them were calculated. The relations between the normalized average ratios of the effective and ineffective siRNAs for the recently reported genes are shown in Figure 2.

Evaluation using nucleotide frequencies based on 833 effective siRNAs

Case 1: Independent nucleotide occurrences at individual positions: The average normalized ratio ***NR*** for the MG1 effective siRNAs was 1.14, whereas that for the ineffective ones was 0.26. This indicates that as the ***NR*** for the sequences of MG1 effective siRNAs are 1.14 times higher than that for the 833 effective siRNAs, it shows the higher level potential in gene-silencing. On the other hand, as the ***NR*** for the sequences of MG1 ineffective ones shows 0.256 times, i.e., one-fourth compared to the ***NR*** for the 833 effective siRNAs, it implies one-fourth (low) level potential of gene-silencing. Since the average normalized ratios for MG2 effective and ineffective siRNAs were respectively 1.06 and 0.25, they indicate a similar tendency of MG1. In contrast, the average normalized ratios for MG3 effective and ineffective siRNAs were respectively 0.82 and 0.46. These results indicate that there is no big difference between them (compared to the MG1 and MG2 effective and ineffective siRNAs). That is, the nucleotide frequency characteristics of MG3 effective siRNAs resemble those of MG3 ineffective siRNAs. Although the ratios of the average effective-to-ineffective ratios for MG1 and MG2 are respectively 4.45 (1.14/0.26) and 4.08 (1.06/0.25), the average effective-to-ineffective ratio for MG3 is 1.78 (0.82/0.46). Since the average normalized ratios of MG4 effective and ineffective siRNAs were respectively 2.14 and 0.13, the ratio of the effective to ineffective siRNAs was 16.5 (2.14/0.13). The ***NR*** of the effective siRNAs for MG4 therefore implies a high likelihood (2.2 times) of gene-silencing compared to that of the 833 effective siRNAs, whereas the ***NR*** of the ineffective ones show quite low likelihood (0.13 times). On the other hand, since the normalized ratios for MG5 were respectively 0.68 and 0.24, the ratio of the effective and ineffective siRNAs was 2.83. The entire normalized ratio that effective siRNAs for MG1 to MG5 would be effective was 1.06, whereas the entire normalized ratio that the ineffective ones would be effective was 0.297. These evaluation results for the independent nucleotide occurrences indicate that the proposed prediction method based on the effective siRNA sequences is useful for selecting candidate siRNAs for target genes.

Case 2: Dependent nucleotide occurrences based on the simple Markov model: As shown in Figure 2, in Case 2 as a whole the average normalized ratios of the effective and ineffective siRNAs for MG1, MG2, MR3, and MG5 were lower than those in Case 1. In contrast, the normalized ratio of MG4 effective siRNAs was higher than that in Case 1 and the normalized ratio of MG4 ineffective ones was lower than that in Case 1. There is, however, a similar tendency in the ratios of the effective–to-ineffective average ratios for MG1, MG2, MG3, and MG5. The average normalized ratio of the MG1 effective siRNAs was 0.79, whereas that of the ineffective ones was 0.19. The average ratio of the effective siRNAs is thus about four times larger than that of the ineffective ones. As the average normalized ratios of the MG2 effective and ineffective siRNAs were respectively 0.59 and 0.04, the average ratio of the effective siRNAs was about 14 times larger than that of the ineffective ones. On the other hand, the average normalized ratios of the MG3 effective and ineffective siRNAs were respectively 0.77 and 0.31. Although the average ratio of the effective siRNAs was only about 2.5 times larger than that of the ineffective ones and this ratio was lower than the corresponding ratios for the MG1 and MG2 siRNAs, there was still a clear difference between the average normalized ratios of the MG3 effective and ineffective siRNAs. Similarly, the normalized ratios of the MG5 effective and ineffective siRNAs were respectively 0.37 and 0.15. Therefore the average ratio of effective siRNAs was approximately 2.5 times larger than that of the ineffective ones. On the other hand, since the average normalized ratios of MG4 (cyclin B1) effective and ineffective siRNAs were respectively 2.88 and 0.08, the difference between them was a remarkably large (36-fold). The ***NR*** of the effective siRNAs for MG4 therefore indicated the higher likelihood of gene-silencing compared to that of the 833 siRNAs, whereas the ***NR*** for the ineffective ones showed the quite low likelihood. These evaluation results for the dependent nucleotide occurrences based on the simple Markov model indicate that the proposed prediction method is useful for selecting candidate siRNAs for target genes (Figure 2).

Evaluation using another large number of known siRNAs

Gene-silencing probabilities were also evaluated using the nucleotide frequencies at individual positions in 636 other effective siRNAs. The independent (Case 1) and dependent (Case 2) nucleotide frequencies at individual positions for other 636 effective probabilities that the effective and ineffective siRNAs would be effective for the reported genes are shown in Figure 3. Although there were ups and downs in ***NRs*** predicted for MG1 to MG5 using either the 833 or 636 effective siRNAs, the total ***NRs*** predicted are similar for both cases. That is, the ***NRs*** based on the 833 effective siRNAs are respectively 1.06 and 0.81 for the independent and dependent cases, and those based on the 636 effective siRNAs are respectively 1.1 and 0.87 for the independent and dependent cases. This implies that the proposed method using the average silencing probabilities could be useful for many other genes (Figure 3).

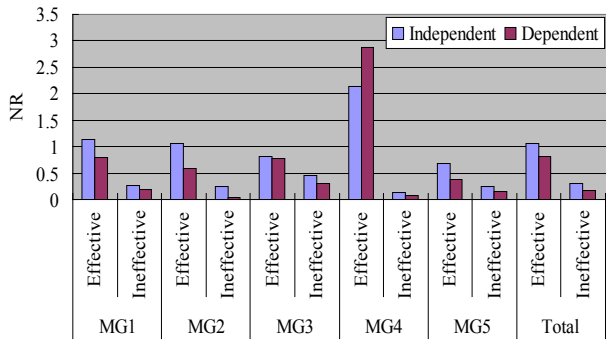

Figure 2: Normalized ratios based on 833 effective siRNAs. Effective: effective siRNAs. Ineffective: ineffective siRNAs. NR: Normalized ratio calculated by equation (10). Independent: independent occurrences at individual positions, Dependent: the simple Markov model

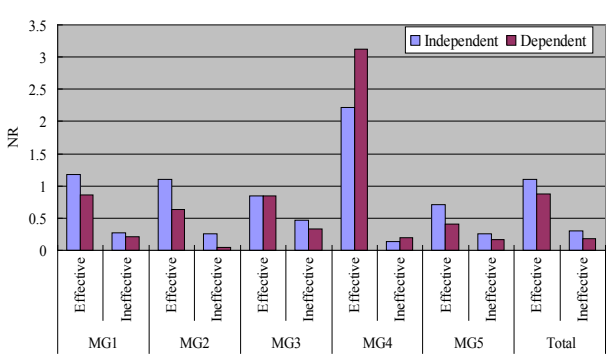

Figure 3: Normalized ratios based on 636 effective siRNAs.

Evaluation for the HMM

The Viterbi algorithm was carried out for the state diagram of the HMM shown in Figure 1. As a result, the siRNA sequence GAAGAAGAGAGAGAGCAGA was obtained as the optimal nucleotide sequence (i.e., the sequence maximizing the sequence state probability for positions 1 to 19). This result also indicates that the nucleotides G and A might dominate the optimal sequence in reported sets of effective siRNAs.

It is also possible to select individual positional nucleotides for minimizing the sequence state probability. This was done by using the modified Viterbi algorithm, i.e., by changing from maximum to minimum in the equations (4) to (9), and yielded the sequence TTTTTATTAATCGCGTTCG. From the point of gene-silencing by siRNA sequences, the optimal maximized sequence may correspond to the most preferable siRNA sequence in a large number of effective siRNAs. On the other hand, the minimized sequence may correspond to the least preferable one in a large number of effective siRNAs.

These maximized and minimized nucleotide sequences were then compared with the upper- and lower-level significant nucleotides obtained using the previously proposed statistical significance testing for 833 effective siRNA sequences [30]. One sees in Table 4 that the maximized nucleotide obtained using the Viterbi algorithm corresponds to the upper-level nucleotides obtained using the significance testing, and the minimized nucleotide sequence corresponds to the lower-level one obtained using the significance testing. Interestingly, there are many coincidences between the maximized and minimized nucleotides and the upper- and lower-level significant nucleotides. Between the maximized nucleotides and the upper-level ones there are thirteen coincidences (at positions 1, 2, 3, 4, 6, 7, 8, 9, 12, 14, 16, 17, and 19), and between the minimized nucleotides and the lower-level ones there are eleven coincidences (at positions 1, 2, 4, 5, 7, 10, 11, 14, 15, 18, and 19). There are six coincidence positions in both relations: at positions 1, 2, 4, 7, 14, and 19. The positions 1, 2, and 19 correspond to around the 5' and 3' terminal points. This implies that these positions play important roles in gene-silencing.

Evaluation for MG1 to MG5 based on a large number of ineffective siRNAs: It is also possible to clarify the probability of how siRNA candidates are effective on the basis of a large number of ineffective siRNAs. 847 known siRNAs were selected as ineffective ones (see Methods and Materials). The probabilities of individual nucleotide occurrence frequencies at individual positions are listed in Table 3c. The relations among ***NR***s of effective and ineffective siRNAs for MG1 to MG5 computed by using the equations (2), (3), and (10) are shown in Figure 4. In the case of using the 847 known ineffective siRNAs, ***NR***s of effective siRNAs for MG1 to MG5 are less than 1, whereas those of ineffective ones are more than 1 as shown in Figure 4. The ***NR*** of the total effective siRNAs is 0.67, whereas that of the ineffective ones is 2.37.

Comparing Figure 4 with Figure 2, it is clear that the corresponding ***NR***s of effective and ineffective siRNAs for MG1 to MG5 are respectively reverse relations. This depends on what set of siRNAs, i.e., 833 or 847 siRNAs, is used. There are differences in the nucleotide occurrence frequencies between both sets of siRNAs as shown in Figure 5. Especially, there are big differences at positions 1 and 19. These results are also useful for designing effective siRNA sequences (Figures 4 and 5).

Comparison with other reported methods

The proposed method uses a probability estimation technique for selecting effective siRNA candidates, whereas most of the previous methods use scoring techniques. Although it is not easy to compare the previously reported scoring methods with the proposed average probability technique, the relations between the ratios of the scores for effective and ineffective siRNAs and the probabilities predicted for effective and ineffective siRNAs by using the average probabilities can be compared by analyzing the ROC (Receiver Operating Characteristic) curve based on the True Positive Fraction (TPF) and the False Positive Fraction (FPF) [44,45]. Because the reliability of the ROC curves increases with the numbers of effective and ineffective siRNAs that are used, 833 and 103 (MG1 – MG5) effective siRNAs and 847 and 93 (MG1 – MG5) ineffective siRNAs were adopted (see Methods and Materials). The ROC curves generated using the proposed method and the previously reported scoring methods are shown in Figure 6, where one sees that the curve for the proposed method is similar to those for the methods of Ui-Tei et al. [19] and Amarzguioui and Prydz [7] and is superior to those for Reynolds et al. [18] and Hsieh et al. [21]. This indicates that the proposed method distinguishes between effective and ineffective siRNAs as well as the previous top-ranked scoring techniques do. Furthermore, the previous scoring techniques cannot estimate the probability that a candidate siRNA with high score will actually accomplish the expected gene degradation, whereas the proposed method can do it (Figure 6).

Conclusions

This paper proposed an analytical prediction method using the

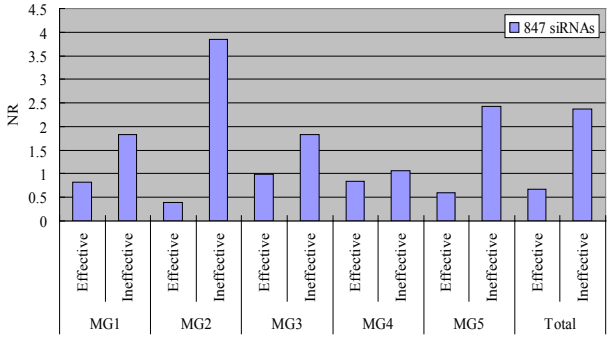

Figure 4: Normarized ratios based on 847 ineffective siRNAs.

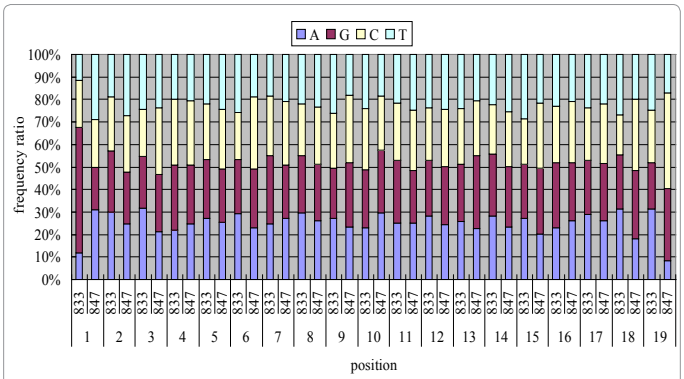

Figure 5: Relations of nucleotide occurrence frequencies between 833 and 847 siRNAs.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Upper level	G	A	A/T	C/G	C	A/T	G/C	A	C/G	C	C	A/C		A	T/A	T/C	A	T/A	A
Maximized	G	A	A	G	A	A	G	A	G	A	G	A	G	A	G	C	A	G	A
Coincidence	=	=	=	=		=	=	=	=			=		=		=	=		=
Lower-level	A/T	T	G	T/A	T	G/C	T/A	C	G	A/G	T			C	G/C	A	G	C/G	G
Minimized	T	T	T	T	T	A	T	T	A	A	T	C	G	C	G	T	T	C	G
Coincidence	=	=		=	=		=			=	=			=	=			=	=

Table 4: Relations between the maximized and minimized nucleotide sequences and the upper- and lower-level significant nucleotides.

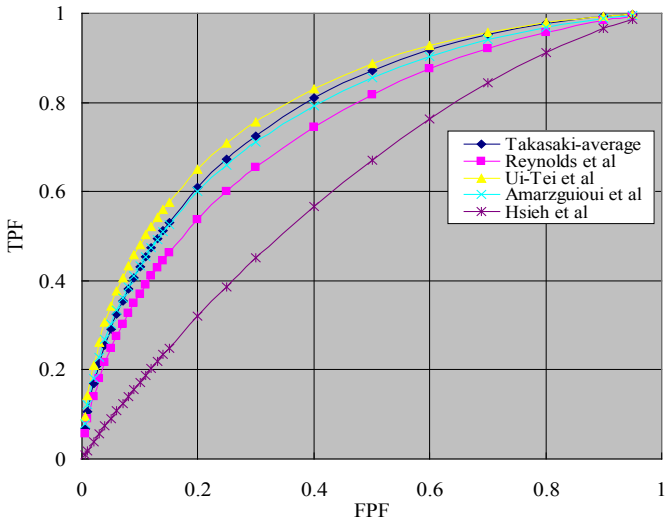

Figure 6: ROC curves comparing the proposed method with previously reported scoring techniques.ROC curves of the individual scoring techniques and the proposed method were generated for 936 effective and 940 ineffective siRNAs[44]. Scores of the effective and ineffective siRNA sequences were computed on the basis of the positional scores of the individual guidelines shown in Saetrom and Snove [36].

average silencing probability to select effective siRNA target sequences from many possible candidate sequences. Although the previous scoring methods cannot estimate the probability that a candidate siRNA sequence will actually accomplish the expected gene degradation, the proposed method can. It is therefore quite different from the previous scoring methods. The proposed method was evaluated by applying it to recently reported siRNA sequences effective and ineffective for various genes. The evaluation results indicate that the proposed method would be useful for many other genes. It should therefore be useful for selecting siRNA sequences for mammalian genes. The paper also described another method using a Hidden Markov Model (HMM) to select the optimal functional siRNAs.

References

1. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, et al. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391: 806-811.
2. Sharp PA (2001) RNA interference--2001. Genes Dev 15: 485-490.
3. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, et al. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411: 494-498.
4. Elbashir SM, Lendeckel W, Tuschl T (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev 15: 188-200.
5. Dykxhoorn DM, Novina CD, Sharp PA (2003) Killing the messenger: short RNAs that silence gene expression. Nat Rev Mol Cell Biol 4: 457-467.
6. Hannon GJ (2002) RNA interference. Nature 418: 244-251.
7. Holen T, Amarzguioui M, Wiiger MT, Babaie E, Prydz H (2002) Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. Nucleic Acids Res 30: 1757-1766.
8. Elbashir SM, Martinez J, Patkaniowska A, Lendeckel W, Tuschl T (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. EMBO J 20: 6877-6888.
9. Kumar R, Conklin DS, Mittal V (2003) High-throughput selection of effective RNAi probes for gene silencing. Genome Res 13: 2333-2340.
10. Mittal V (2004) Improving the efficiency of RNA interference in mammals. Nat Rev Genet 5: 355-365.
11. Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, et al. (2003) Asymmetry in the assembly of the RNAi enzyme complex. Cell 115: 199-208.
12. Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. Cell 115: 209-216.
13. Chalk AM, Wahlestedt C, Sonnhammer EL (2004) Improved and automated prediction of effective siRNA. Biochem Biophys Res Commun 319: 264-274.
14. Teramoto R, Aoki M, Kimura T, Kanaoka M (2005) Prediction of siRNA functionality using generalized string kernel and support vector machine. FEBS Lett 579: 2878-2882.
15. Naito Y, Yamada T, Ui-Tei K, Morishita S, Saigo K (2004) siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. Nucleic Acids Res 32: W124-129.
16. Santoyo J, Vaquerizas JM, Dopazo J (2005) Highly specific and accurate selection of siRNAs for high-throughput functional assays. Bioinformatics 21: 1376-1382.
17. Truss M, Swat M, Kielbasa SM, Schäfer R, Herzel H, et al. (2005) HuSiDa--the human siRNA database: an open-access database for published functional siRNA sequences and technical details of efficient transfer into recipient cells. Nucleic Acids Res 33: D108-D111.
18. Reynolds A, Leake D, Boese Q, Scaringe S, Marshall WS, et al. (2004) Rational siRNA design for RNA interference. Nat Biotechnol 22: 326-330.

19. Ui-Tei K, Naito Y, Takahashi F, Haraguchi T, Ohki-Hamazaki H, et al. (2004) Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Res 32: 936-948.

20. Jiang P, Wu H, Da Y, Sang F, Wei J, et al. (2007) RFRCDB-siRNA: improved design of siRNAs by random forest regression model coupled with database searching. Comput Methods Programs Biomed 87: 230-238.

21. Hsieh AC, Bo R, Manola J, Vazquez F, Bare O, et al. (2004) A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. Nucleic Acids Res 32: 893-901.

22. Jagla B, Aulner N, Kelly PD, Song D, Volchuk A, et al. (2005) Sequence characteristics of functional siRNAs. RNA 11: 864-872.

23. Huesken D, Lange J, Mickanin C, Weiler J, Asselbergs F, et al. (2005) Design of a genome-wide siRNA library using an artificial neural network. Nat Biotechnol 23: 995-1001.

24. Snove O Jr, Nedland M, Fjeldstad SH, Humberset H, Birkeland OR, et al. (2004) Designing effective siRNAs with off-target control. Biochem Biophys Res Commun 325: 769-773.

25. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis – probabilistic models of proteins and nucleic acids, Cambridge University Press, UK.

26. Elbashir SM, Harborth J, Weber K, Tuschl T (2002) Analysis of gene function in somatic mammalian cells using small interfering RNAs. Methods 26: 199-213.

27. Holen T (2006) Efficient prediction of siRNAs with siRNArules 1.0: an open-source JAVA approach to siRNA algorithms. RNA 12: 1620-1625.

28. Saetrom P (2004) Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. Bioinformatics 20: 3055-3063.

29. Ladunga I (2007) More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. Nucleic Acids Res 35: 433-440.

30. Takasaki S, Kawamura Y, Konagaya A (2006) Selecting effective siRNA sequences based on the self-organizing map and statistical techniques. Comput Biol Chem 30: 169-178.

31. Takasaki S, Konagaya A (2006) Comparative analyses for selecting effective siRNA sequences. Chem-Bio Informatics J 6: 69-84.

32. Takasaki S, Kawamura Y (2007) Using radial basis function networks and significance testing to select effective siRNA sequences. Computational Statistics & Data Analysis 51: 6476-6487.

33. Heale BS, Soifer HS, Bowers C, Rossi JJ (2005) siRNA target site secondary structure predictions using local stable substructures. Nucleic Acids Res 33: e30.

34. Luo KQ, Chang DC (2004) The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. Biochem Biophys Res Commun 318: 303-310.

35. Bohula EA, Salisbury AJ, Sohail M, Playford MP, Riedemann J, et al. (2003) The efficacy of small interfering RNAs targeted to the type 1 insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript. J Biol Chem 278: 15991-15997.

36. Saetrom P, Snøve O Jr (2004) A comparison of siRNA efficacy predictors. Biochem Biophys Res Commun 321: 247-253.

37. Shabalina SA, Spiridonov AN, Ogurtsov AY (2006) Computational models with thermodynamic and composition features improve siRNA design. BMC Bioinformatics 7: 65.

38. Vert JP, Foveau N, Lajaunie C, Vandenbrouck Y (2006) An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinformatics 7: 520.

39. Lu ZJ, Mathews DH (2008) Efficient siRNA selection using hybridization thermodynamics. Nucleic Acids Res 36: 640-647.

40. Wang X, Wang X, Varma RK, Beauchamp L, Magdaleno S, et al. (2009) Selection of hyperfunctional siRNAs with improved potency and specificity. Nucleic Acids Res 37: e152.

41. Takasaki S, Kawamura Y, Konagaya A (2006) Selecting effective siRNA sequences by using radial basis function network and decision tree learning. BMC Bioinformatics 7: S22.

42. Takasaki S, Kotani S, Konagaya A (2004) An effective method for selecting siRNA target sequences in mammalian cells. Cell Cycle 3: 790-795.

43. Takasaki S, Kotani S, Konagaya A (2005) Selecting effective siRNA target sequences for mammalian genes. RNA Biol 2: 21-27.

44. Metz CE, Herman BA, Roe CA (1998) Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. Med Decis Making 18: 110-121.

45. Jensen FV, Nielsen TD (2001) Bayesian Networks and Decision Graphs, Springer.