

Chapter 2

Methods for Selecting Effective siRNA Target Sequences Using a Variety of Statistical and Analytical Techniques

Shigeru Takasaki

Abstract

Short interfering RNA (siRNA) has been widely used for studying gene function 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 chapter first reviews the recently reported siRNA design guidelines and then proposes new methods for selecting effective siRNA sequences from many possible candidates by using decision tree learning, Bayes' theorem, and average silencing probability on the basis of a large number of known effective siRNAs. These methods differ from the previous score-based siRNA design techniques and can predict the probability that a candidate siRNA sequence will be effective. Evaluation of these methods by applying them to recently reported effective and ineffective siRNA sequences for a number of genes indicates that they would be useful for many other genes. They should, therefore, be of general utility for selecting effective siRNA sequences for mammalian genes. The chapter also describes another method using a hidden Markov model to select the optimal functional siRNAs and discusses the frequencies of combinations of two successive nucleotides as an important characteristic of effective siRNA sequences.

Key words: siRNA, siRNA design, RNA interference, Gene silencing, Estimation of gene silencing, Decision tree learning, Bayes' theorem, Average gene silencing, Hidden Markov model

1. 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 short interfering RNA

(siRNA) sequences for target genes. The effectiveness of the 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. showed that the 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 (11, 12). 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 (18–23). This implies that they might result in the generation of many candidate target sequences, making it difficult to select the effective ones. In addition, the previously reported rules 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 those 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 chapter first reviews the recently reported siRNA design guidelines and clarifies their problems. It then describes prediction methods for selecting effective siRNA target sequence from many possible candidate sequences by using decision tree learning, Bayes' theorem, and average silencing probability of a large number of siRNA sequences known to be effective (25–32). They are 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 these statistical methods to recently reported effective and ineffective siRNA sequences for various genes showed that they are accurate and thus imply that they would be useful for selecting siRNA sequences silencing many other genes. This chapter also describes another method using a hidden Markov model (HMM) to select the optimal functional siRNAs (25) and discusses the frequencies of combinations of two successive nucleotides as an important characteristic of effective siRNA sequences.

2. 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. They have so far been identified by using a trial-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 trial-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.

2.1. The Reported Guidelines for Designing siRNA Sequences

The earliest guidelines for siRNA sequence design were proposed by Elbashir et al. (4, 8, 33). 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 chapter considers the following five (herein 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 three 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 (20) analyzed 46 siRNAs targeting four genes and reported six rules for effective siRNA designs.

Table 1
Effective and ineffective nucleotides specified
in the individual guidelines

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

Position: Nucleotide position from 1 to 19 (5' to 3', cDNA form)
Preferred: Effective (positive), unpreferred: ineffective (negative)

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 , and (7) energy difference between -1 and 0 . The score of an siRNA candidate is incremented by one point for each rule fulfilled and is thus between 0 and 7 .

Huesken et al. (23) reported a method for screening functional siRNAs by using an artificial neural network. This network was first trained by 2,182 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 (34) have reported functional siRNA selection methods using support vector machines (SVMs). Teramoto et al. used a generalized string kernel (GSK) combined with an SVM. siRNA sequences were represented as vectors in a multidimensional feature space according to the number of subsequences in each siRNA and were classified as effective or ineffective (14). Ladunga used an SVM with polynomial kernels and constrained optimization models from 572 sequence, thermodynamic, accessibility, and self-hairpin features over 2,200 published siRNAs (23, 34). 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 recently reported siRNA rules based on apparent overrepresentation or underrepresentation of certain nucleotides in certain positions of the Novartis data set (35). The criteria for an siRNA candidate depend on the positive and negative scores computed for each position by using a 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, the original scores in the scoring table are based on the percentage overrepresentation or underrepresentation of certain nucleotides in certain positions and thus 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 (36–38), 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 (39, 40).

Table 2
Features of individual siRNA design rules/algorithms

siRNA design rules	Citation	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	2,128	Sequence motifs	Neural network
Jagla et al.	(22)	4	601	Sequence features	Decision tree
Holen	(34)	34	400	Sequence features	Percentage
Saetrom	(35)	40	581	Sequence motifs	Genetic programming
Teramoto et al.	(14)	2	94	Sequence motifs	Support vector machine
Ladunga	(41)	34	2,252	Position features	Support vector machine
Chalk et al.	(13)	92	398	Binding energy	Regression tree
Takasaki et al.	(29–31)	490	833	Sequence features	Statistics, SOM

The features of various siRNA design rules are summarized in Table 2. In addition, other design rules using a combination of the above techniques have also been used to obtain efficacious siRNAs based on public/open siRNA data sets (42–54).

2.2. Problems with the Previous Guidelines

Among the problems with the reported guidelines is the problem of inconsistencies with regard to the nucleotide frequencies of 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 identifying effective sequences for some genes, they might be unsuitable for others. Because 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 siRNA target sequence candidates for other mammalian genes, many ineffective sequences might be selected as candidates. This is due to the prevalence of long coding regions in mammalian genes. This poses an additional problem because experimentally evaluating whether the selected sequences provide effective gene degradation is a costly and time-consuming task. 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-scoring siRNA were obtained using the

Table 3
Relations between attributes (positions and nucleotides of 19 nt sequences)
and training instances (no. of effective and ineffective siRNAs)

Position	1				2				3				⋮	19			
Nucleotide	A	G	C	T	A	G	C	T	A	G	C	T	⋮	A	G	C	T
No. of nt (effect)	100	464	173	96	250	225	202	156	265	191	173	204	⋮	259	173	194	207
No. of nt (ineffect)	264	157	182	244	209	194	214	230	179	217	251	200	⋮	71	270	361	145

No. of nt (effect): No. of nucleotides in effective siRNAs

No. of nt (ineffect): No. of nucleotides in ineffective siRNAs

reported methods, it would be difficult to estimate the probability that it would actually accomplish the expected gene degradation. To overcome the problems of the previous guidelines, Takasaki et al. recently reported new scoring methods using the statistical and clustering techniques listed in Table 2 (28–32).

3. Methods for siRNA Sequence Selections

To design effective siRNA sequences for target genes, decision tree learning, Bayes's theorem, and average silencing probability of a large number of effective siRNA sequences were used in the proposed methods (26, 27). These methods mainly consist of two phases, one is to learn the relations between individual siRNA sequences and their gene silencing efficacies by using known data, and the other is to predict gene silencing probabilities for new candidate siRNA sequences by using the learned relations. The learning phase is carried out by training and validation phases by supplying many known effective and ineffective siRNA sequences. The prediction phase is where predictions of gene silencing probabilities are actually computed for new candidate siRNA sequences.

3.1. Prediction by the Decision Tree Learning Method

3.1.1. Model for Decision Tree Learning

Individual positions and nucleotides of 19 sequences listed in Table 3 were used as attributes. 833 effective and 847 ineffective siRNA sequences were used as training instances.

To carry out the supervised learning for effective siRNA classifications by using decision tree learning, the training instances are partitioned into two sets, one for the growth of the decision tree (training data) and other for the decision tree pruning (testing data). The processes of the classifications were carried out in two phases: the growth and pruning of the decision tree.

3.1.2. The Growth of the Decision Tree

The algorithm, in outline, is as follows:

1. If all the instances belong to a single class, there is nothing to do (except create a leaf node labeled with the name of that class).
2. Otherwise, for each attribute that has not already been used, calculate the information gain that would be obtained by using that attribute on the particular set of instances classified to this branch node. The information gain can be computed in the following way (55):

$$I(p, u) = -\frac{p}{p+u} \log_2 \left(\frac{p}{p+u} \right) - \frac{u}{p+u} \log_2 \left(\frac{u}{p+u} \right) \quad (1)$$

where p is the total number of nucleotides for this attribute in effective (preferred) siRNA sequences and u is the total number of nucleotides in ineffective (unpreferred) siRNA sequences.

The entropy $H(L)$ associated with the attribute L is

$$H(L) = \sum_{i=1}^v \frac{p_i + u_i}{p + u} I(p_i, u_i) \quad (2)$$

where v is a kind of nucleotide, i.e., $i=1=A$, $2=G$, $3=C$, and $4=T$, p_i and u_i are, respectively, a number of the corresponding nucleotides in effective and ineffective siRNA sequences, and L is the attribute (position) listed in Table 3. The information gain is therefore obtained as follows:

$$\text{gain}(L) = I(p, u) - H(L) \quad (3)$$

3. Use the attribute (position) with the greatest information gain as a branch node and for each nucleotide of L , create a new descendant of the node.
4. If the information gain becomes less than the specified criterion, stop the growth of the decision tree and create leaf nodes; otherwise continue to build the tree.

3.1.3. Decision Tree Pruning

Working backwards from the bottom of the tree, the subtree starting at each nonterminal node is examined. If the error (misclassification) rate on the testing data improves by pruning it, the subtree is removed. The process continues until no improvement can be made by pruning a subtree.

The predictions of gene silencing for new candidate siRNA sequences are carried out by the decision tree method described above. This article used the recently reported effective and ineffective siRNA sequences as the evaluation data for predictions (see Subheading 3.5).

3.2. Prediction Analysis of siRNA Target Sequences Based on Bayes' Theorem

Bayes' theorem tells us the probability that a hypothesis is true given the prior probability that we would have assigned to the hypothesis, and both conditional probabilities A and B. In this case, the hypothesis is that an siRNA candidate will effectively silence mammalian genes, the prior probability is the best guess,

i.e., the empirical information, the conditional probability A is that an siRNA candidate will effectively silence genes if the individual nucleotides at each of the positions of the candidate belonged to the set of nucleotide frequently found in effective siRNAs, and the conditional probability B is that the siRNA candidate will effectively silence genes even if the corresponding nucleotides belonged to the set of nucleotides frequently found in ineffective siRNAs. The sets of nucleotide frequencies for effective and ineffective siRNAs can be estimated from a large number of known effective and ineffective siRNA sequences in the literature.

3.2.1. Prediction of Gene Degradation Ratio for a Given siRNA Candidate

Given a candidate siRNA sequence \mathbf{X} ($\mathbf{X} = X_1, X_2, \dots, X_{19}$, where X_i is a nucleotide) for a specified target gene, Bayes' theorem (25) could be used to predict the following gene silencing probability $P(\text{eff}|\mathbf{X})$ for that sequence:

$$P(\text{eff}|\mathbf{X}) = \frac{P^{\text{eff}} P(\mathbf{X} | \text{eff})}{P(\mathbf{X})} \quad (4)$$

where P^{eff} is the prior probability of more than 80% gene reduction by siRNA sequences. This probability is obtained, as empirical knowledge, from many siRNA experiments and is known to be approximately 0.1–0.2 in mammalian genes (9, 10, 18). This prior probability represents the best guess that we can make about an siRNA candidate sequence before we have seen information about the sequence itself. $P(\mathbf{X}|\text{eff})$ is the probability that \mathbf{X} would cause effective gene silencing if X_1, X_2, \dots, X_{19} belonged to the set of nucleotides frequently found in effective siRNA sequences, for example, from 833 effective siRNAs in the literature (see Subheading 3.5.1). $P(\mathbf{X}|\text{eff})$ is therefore computed by the product of individual frequency ratios of the positional nucleotides in the following way:

$$P(\mathbf{X}|\text{eff}) = \prod_{i=1}^{19} q_{x_i^n}^{\text{eff}} \quad (5)$$

where i is the nucleotide position and n is the kind of nucleotide (A, C, G, or T). That is, $x_i^A = A$, $x_i^C = C$, $x_i^G = G$, and $x_i^T = T$ and $q_{x_i^n}^{\text{eff}}$ is the frequency ratio of i th position nucleotide n of the candidate \mathbf{X} . Individual $q_{x_i^n}^{\text{eff}}$ can be derived from the corresponding nucleotide frequency ratios in the set of effective siRNA sequences described before. As $P(\mathbf{X})$ is the probability that \mathbf{X} will result in effective gene silencing, it can be computed by summation of both probabilities derived from frequency ratio sets of known effective and not-effective siRNAs in the following way:

$$P(\mathbf{X}) = P^{\text{eff}} P(\mathbf{X}|\text{eff}) + P^{\text{inf}} P(\mathbf{X}|\text{inf}) \quad (6)$$

where P^{inf} is also the prior probability and is equal to $1 - P^{\text{eff}}$.

$P(\mathbf{X}|\text{inf})$ is the probability that \mathbf{X} will result in effective gene silencing if X_1, X_2, \dots, X_{19} belongs to the set of nucleotides frequently

found in not-effective sequences. $P(X|\text{inf})$ is therefore computed by the product of individual frequency ratios of the positional nucleotides in the following way:

$$P(X|\text{inf}) = \prod_{i=1}^{19} q_{x_i^n}^{\text{inf}} \quad (7)$$

where $q_{x_i^n}^{\text{inf}}$ is the frequency ratio of i th position nucleotide n (A, C, G, or T) of the candidate X . Individual $q_{x_i^n}^{\text{inf}}$ can be obtained from the corresponding nucleotide frequency ratios in the set of ineffective siRNA sequences (see Subheading 3.5.2).

Using Eqs. 5–7, we can express Eq. 4 as follows:

$$P(\text{eff}|X) = \frac{P^{\text{eff}} P(X|\text{eff})}{P^{\text{eff}} P(X|\text{eff}) + P^{\text{inf}} P(X|\text{inf})} = \frac{P^{\text{eff}} \prod_{i=1}^{19} q_{x_i^n}^{\text{eff}}}{P^{\text{eff}} \prod_{i=1}^{19} q_{x_i^n}^{\text{eff}} + P^{\text{inf}} \prod_{i=1}^{19} q_{x_i^n}^{\text{inf}}} \quad (8)$$

This $P(\text{eff}|X)$ is the gene silencing probability we want. It is called the posterior probability that X will result in gene silencing because it is our best guess after we have seen the siRNA candidate sequence.

Suppose, for example, that we have a candidate siRNA sequence CCATCAACACCGAGTTCAA for some target gene. What is the probability of gene silencing by this sequence? If we assume that $p^{\text{eff}}=0.1$ and $p^{\text{inf}}=0.9$ and the observed frequencies of effective ($q_{x_i^n}^{\text{eff}}$) and ineffective ($q_{x_i^n}^{\text{inf}}$) nucleotides at each of the positions are given as $q_{x_i^n}^{\text{eff}} = (0.208, 0.242, 0.318, 0.197, 0.248, 0.294, 0.247, 0.229, 0.271,$

$0.275, 0.256, 0.247, 0.259, 0.276, 0.286, 0.23, 0.235, 0.313, 0.312)$ and

$q_{x_i^n}^{\text{inf}} = (0.215, 0.253, 0.211, 0.207, 0.266, 0.231, 0.273, 0.253, 0.235, 0.242, 0.269, 0.257, 0.226, 0.266, 0.215, 0.21, 0.262, 0.182, 0.084)$ from the sets of nucleotide frequencies in the effective and ineffective siRNA sequences, Eq. 8 predicts that the gene silencing probability of this candidate is 0.63 (63%). This indicates that although we have only information about 10% gene silencing by this siRNA candidate as the prior probability (the best guess) before we have seen the candidate, we can predict a gene silencing probability six times higher after we have seen it. This is useful information for selecting effective siRNA candidates. In the case of $p^{\text{eff}}=0.2$ and $p^{\text{inf}}=0.8$, the gene silencing probability of the candidate is predicted to be 79.3%.

3.2.2. Evaluation Criteria Used in the Proposed Method

To make the estimated results using Bayes' theorem easily understood, the ratio of the result estimated for a new siRNA candidate to the prior probability (best guess: empirical information) is first considered. This ratio PR is therefore defined as follows:

$$PR = \frac{ER}{PP} \quad (9)$$

where PP is the prior probability and ER is the result estimated by Bayes' theorem.

If $PR > 1$, it indicates that the level of gene silencing will be higher than the usual level of silencing. Conversely, a $PR < 1$ indicates that the level of silencing will be lower than the usual level.

Now let us consider the normalized ratio (NR) obtained when we divide the result estimated for a new siRNA candidate by the average of the results estimated for a large number of known effective siRNAs. We do this because the results estimated for the known effective siRNAs could be considered a standard criterion for new siRNA candidates. The ratio NR is therefore defined as follows:

$$NR = \frac{ER}{AP} \quad (10)$$

where AP 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 $NR \geq 1$, the level of gene silencing expected to be obtained with the siRNA candidate is the same as or higher than the level of silencing obtained with the known effective siRNAs. That is, NR indicates that the candidate sequence has a high potential for gene silencing. If, on the other hand, $NR < 1$, it indicates that the gene silencing expected to be obtained with the candidate sequence is lower than the level of silencing obtained with the known effective siRNAs.

3.2.3. Verification Models of the Proposed Method

We can see from Eq. 8 that the accuracy of the gene silencing probability predicted by the proposed method is greatly dependent on individual $q_{x_i^n}^{\text{eff}}$ and $q_{x_i^n}^{\text{inf}}$ values. Generally, the larger the sets of effective and ineffective siRNAs, the higher the prediction accuracy, and a set of $q_{x_i^n}^{\text{inf}}$ would be a complementary set of $q_{x_i^n}^{\text{eff}}$. It is difficult, however, to obtain complete sets of effective and ineffective siRNAs because of the difficulty of the siRNA verification for all genes. Individual $q_{x_i^n}^{\text{eff}}$ and $q_{x_i^n}^{\text{inf}}$ values were therefore generated on the basis of a large number of effective and ineffective siRNAs in the literature. We make two assumptions about the occurrence of nucleotides in the sets of effective siRNA sequences from which $q_{x_i^n}^{\text{eff}}$ is determined. One is that the nucleotides occur independently at individual positions (see Table 4a), 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 are dependent on the nucleotides at the previous position. The probability of the simple Markov model $M(q_{x_i^n}^{\text{eff}})$ for the nucleotide x_i^n at the present position i is therefore expressed as follows:

$$M(q_{x_i^n}^{\text{eff}}) = P(x_i^n | q_{x_{i-1}^n}^{\text{eff}}) \quad (11)$$

Table 4

Probabilities of individual nucleotide occurrences at each position

(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	0.2229	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		0.203

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	0.235	0.231	0.235
	0.356		0.284	0.314	0.329		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		0.19		0.251		0.196		0.259		0.254			0.235		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		0.274		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		0.16		0.278		0.151		0.23		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		0.146		0.16		0.107		0.148		0.133			0.108		
	0.262		0.239	0.241	0.214		0.214		0.204		0.187		0.167		0.204			0.23		
	0.301		0.3	0.308	0.316		0.316		0.304		0.231		0.263		0.306			0.23		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		0.357		0.262		0.396		0.307		0.325			0.272		
	0.216		0.267	0.281	0.226		0.226		0.273		0.236		0.313		0.173			0.263		
	0.256		0.267	0.194	0.236		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	0.43	0.204	0.255	0.215	0.21	0.221	0.198	0.171

(continued)

Table 4
(continued)

(d) Probabilities of deductive nucleotide occurrences from 833 effective siRNAs

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	0.293	0.233	0.227	0.261	0.243	0.235	0.251	0.234	0.243	0.257	0.25	0.239	0.247	0.239	0.243	0.256	0.237	0.229	0.229
G	0.148	0.243	0.257	0.236	0.246	0.254	0.232	0.249	0.259	0.248	0.24	0.251	0.248	0.241	0.253	0.238	0.253	0.254	0.264
C	0.264	0.253	0.264	0.235	0.251	0.264	0.246	0.257	0.252	0.242	0.248	0.255	0.251	0.261	0.266	0.25	0.255	0.274	0.256
T	0.295	0.271	0.252	0.268	0.261	0.247	0.271	0.26	0.246	0.254	0.261	0.255	0.254	0.259	0.238	0.257	0.255	0.243	0.251

Equation 2.11 tells us that the probability of the nucleotide occurrence x_i^n at the present position i is determined under the condition of the effective nucleotide $q_{x_{i-1}^n}^{\text{eff}}$ at the previous position $i-1$. Suppose, for example, that we have a sequence CCATCAACACCGAGTTCAA as a candidate siRNA sequence for some target gene. In this case, the second nucleotide “C” may occur depending on the first nucleotide “C.” From Table 4b (the simple Markov model table) one sees that the probability that C occurs at position 2 under the condition that there is a 0.208 probability that C occurs at position 1 is 0.295. One similarly sees that the probability that the third nucleotide is A is 0.371 under the condition that there is a 0.242 probability that the second nucleotide is C. Two types of evaluations were therefore considered for the proposed prediction method: (a) one for independent nucleotide occurrences at individual positions and (b) one for nucleotide occurrences dependent on the previous nucleotide occurrences at individual positions.

The set from which $q_{x_i^n}^{\text{inf}}$ is determined could also be assumed to have two types of nucleotide occurrences: independent nucleotide occurrences at individual positions (see Table 4c) and the deductive nucleotide occurrences induced from the effective siRNAs. That is, individual nucleotides could occur as deductive complement ratios of the effective siRNAs. The probabilities of deductive nucleotide occurrences $D(q_{x_i^n}^{\text{inf}})$ at the position i are computed as follows:

$$D(q_{x_i^n}^{\text{inf}}) = \frac{(1 - q_{x_i^n}^{\text{eff}})}{3} \quad (12)$$

Suppose, for example, the deductive probability of the nucleotide “A” occurrence at the position 3. As $q_{x_3^A}^{\text{eff}}$ is 0.318 from Table 4a, $D(q_{x_3^A}^{\text{inf}})$ can be computed as 0.227 by using Eq. 12. The deductive probabilities of individual nucleotides at positions 1–19 are shown in Table 4d. Comparing Table 4a, d, it is clear that the probabilities of individual nucleotide occurrences shown in Table 4d are deductive complement ratios of the effective nucleotides shown in Table 4a. Consequently there could be two cases: (c) one with independent nucleotide occurrences and (d) one with the deductive nucleotide occurrences. The evaluation of the proposed method was therefore carried out in four groups: the combinations of the sets of (a) and (c), (a) and (d), (b) and (c), and (b) and (d). They are, respectively, listed in Table 5 as Cases 1, 2, 3, and 4. Individual nucleotide frequencies at each of the positions of (a), (b), (c), and (d) were generated on the basis of 833 effective siRNA sequences and 847 ineffective siRNA sequences. They are listed in Table 4a–d.

Table 5
The combinations of the proposed method evaluations

	(a) Independent	(b) Dependent
(c) Independent	Case 1	Case 3
(d) Deductive	Case 2	Case 4

Independent: Independent nucleotide occurrences at individual positions

Dependent: Nucleotide occurrences dependent on the simple Markov model

Deductive: Deductive nucleotide occurrences from effective siRNAs

*3.2.4. A Procedure
for Selecting Useful
siRNA Sequences Using
Bayes' Theorem*

Useful siRNA sequences could be selected in the following way.

Estimate the gene silencing potential of candidate siRNA sequences.

- Carry out gene silencing predictions using Bayes' theorem based on the four groups listed in Table 5 (or carry out only the Case 3 because it yielded the most accurate predictions in the evaluations described later).
- Select the siRNA candidates whose gene silencing probability is predicted to be more than 40% when the prior probability is 10%.
- Compare candidate sequences with the sequences of known effective siRNAs.
- Carry out the normalized analyses of the siRNA candidates by using a large number of known siRNAs.
- Select the candidates having a normalized ratio >1.

(Users can easily utilize the proposed method by using Eq. 8 and Table 4.)

**3.3. Procedure for
Selecting Effective
siRNAs 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 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:

$$f_p^N = \frac{\sum_{i=1}^I \{A, G, C, T\}}{I} \quad (13)$$

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 OF_i of each effective siRNA sequence is calculated in the following way:

$$OF_i = \prod_{p=1}^{19} f_{ip}^N \quad (14)$$

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

The average sequence probability A_E for the effective siRNAs is therefore computed as follows:

$$A_E = \frac{\sum_{i=1}^I OF_i}{I} \quad (15)$$

A_E could be considered a criterion for candidate siRNA. That is, if the probability of the candidate sequence were greater than 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 A_E , it would indicate a low likelihood of effectiveness.

3.4. siRNA Sequence Selection Based on a Hidden Markov Model

As an siRNA sequence X basically consists of 19 nucleotides, it can be described as $X = X_1, X_2, \dots, X_{19}$, where X_i indicates the nucleotide A, C, G, or T at position i . Furthermore, this sequence can be expressed as state diagrams of nucleotides A, C, G, and T from the positions 1 to 19 shown in Fig. 1. As shown in Fig. 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 Fig. 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 Fig. 1. Therefore, the transmission of the individual nucleotides A, C, G, 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 an HMM, the optimal states (nucleotides) for maximizing the state transition probability could be solved as a decoding problem by using the Viterbi algorithm (25).

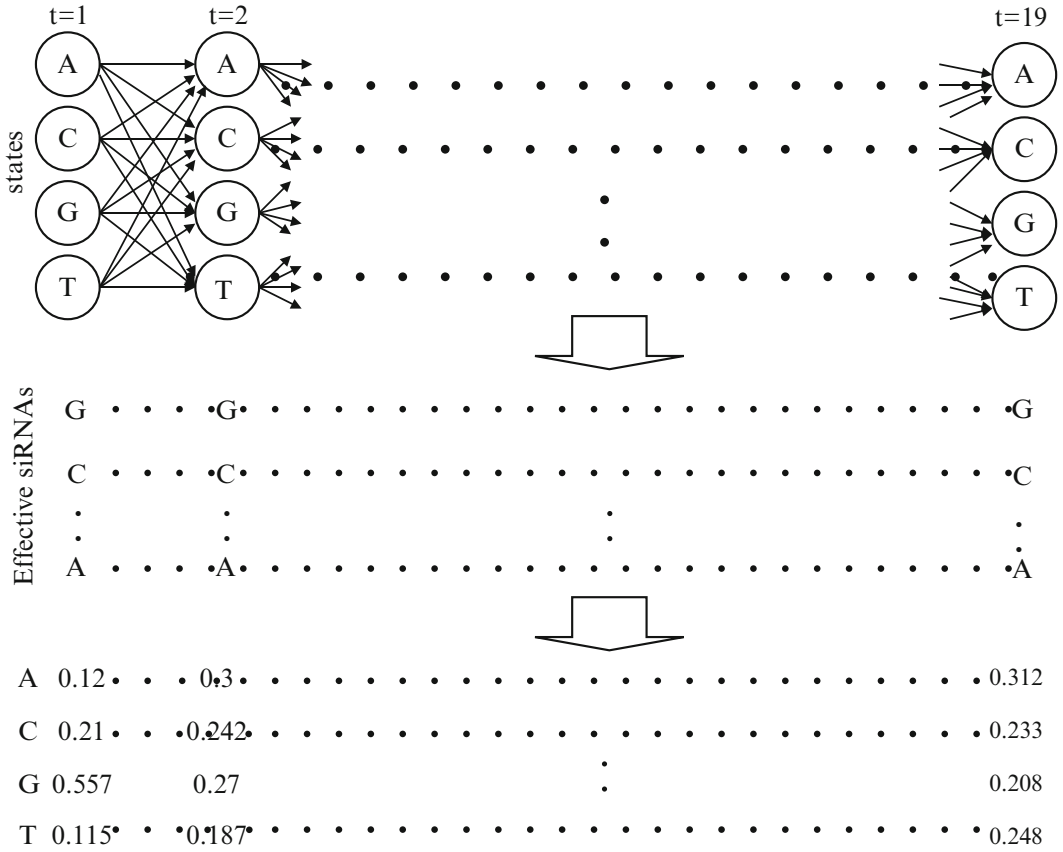


Fig. 1. State diagram of the hidden Markov model. A set of 833 effective siRNA sequences from the literature is shown in the *middle part*, and the frequency ratios of individual nucleotides at each position in those sequences are shown in the *bottom*. The ratios of the nucleotides A, C, G, and T at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

3.4.1. The 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 $i=A, C, G, T$.

$$\delta_1 = \prod_i b_i(O_1) \quad (16)$$

$$\varphi_1(i) = 0,$$

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

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

$$\delta_{t+1}(j) = \max_i (\delta_t(i) a_{ij}) b_j(O_{t+1}) \quad (17)$$

$$\varphi_{t+1}(j) = \arg \max_i (\delta_t(i) a_{ij}) \quad (18)$$

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

3. Termination of the recursive computations.

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

$$\hat{q}_{19} = \arg \max_i \delta_{19}(i) \quad (20)$$

4. Optimal state generation for sequence positions $t=18, 17, \dots, 1$.

$$\hat{q}_t = \varphi_{t+1}(\hat{q}_{t+1}) \quad (21)$$

3.4.2. 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 4a, 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 ($i=2, 3, \dots, 19$) is determined under the condition of the effective nucleotide at the previous position $i-1$ as listed in Table 4b. Suppose, for example, that we have the sequence GACTCAACACCGAGTTCAA as a candidate siRNA sequence for some target gene. In this case, the probability of the second nucleotide being A may depend on the first nucleotide being G. From Table 4b (the simple Markov model table) one sees that the probability that A occurs at position 2 under the condition that there is a 0.557 probability that G occurs at position 1 is 0.33. One similarly sees that the probability that the third nucleotide is C is 0.192 under the condition that there is a 0.3 probability that the second nucleotide is A.

3.4.3. 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 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:

$$NR = \frac{ER}{A_E}, \quad (22)$$

where ER is the result estimated by the average silencing probability method and 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 $NR \geq 1$, the level of gene silencing expected to be obtained with the siRNA candidate is the same as or higher than the 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.

3.5. Training and Testing Data

3.5.1 Effective siRNA Sequences

Two kinds of known effective siRNA sequences were used for the training data. One kind was 833 effective siRNA sequences (more than 80% effective at gene silencing at the protein level) from 490 different cDNAs in the published references of the PubMed database (5, 6, 29). The other was the 636 top-ranked effective siRNA sequences (more than 0.832 (normalized inhibitory activity)) from 34 genes (23).

3.5.2. Ineffective siRNA Sequences

Because ineffective siRNAs are rarely published, we used the 847 worst-ranked ineffective siRNAs (<0.612 (normalized inhibitory activity)) from Huesken et al. (23) for the training data of the proposed methods.

3.5.3. Testing Data

The recently reported effective and ineffective siRNAs were used as the testing data. These testing data were not included in the training data.

Reynolds et al. analyzed 90 siRNAs systematically, targeting every other position of 197-base regions of human *cyclophilin B* mRNA (GeneBank accession no. M60875) (18). For simplicity, human *cyclophilin B* is symbolized throughout the present article as MG1. From the 90 analyzed siRNA sequences we selected as effective ones the 25 top-ranked sequences for which the MG1 target gene silencing was $>80\%$ and selected as ineffective ones the 25 worst-ranked sequences for which MG1 target gene silencing was $<50\%$ with the standard deviation 10%.

Ui-Tei et al. reported 38 effective and 24 ineffective sequences for six genes: *firefly luciferase* (PRL-TK), *vimentin*, *Oct 4*, *EGFP*, *ECFP*, and *DsRed* (19). For simplicity, in the rest of this article all six of these genes are symbolized as MG2.

Amarzguioui and Prydz reported 21 effective and 25 ineffective siRNA sequences for four genes: *hTF* (accession no. M16553), *mTF* (accession no. M26071), *PSK* (accession no. J272212), and *CSK* (accession no. NM_004383) (20). For simplicity, in the rest of this article these four genes are symbolized as MG3.

Takasaki et al. reported seven effective and seven ineffective siRNA sequences for the *cyclin B1* (accession no. NM_031966) (28). For simplicity, in the rest of this article this gene is symbolized as MG4.

Huesken et al. reported 37 siRNAs for *TC10* (accession no. BD135193), *UBE2I* (accession no. NM_003345), and *CDC34* (accession no. NM_004359) (23). The 12 top-ranked effective and 12 worst-ranked ineffective siRNA sequences were selected for these genes. For simplicity, they are symbolized as MG5 in the rest of this article.

4. Evaluations of the Proposed Methods

4.1. Evaluation of the Decision Tree Learning Method

The decision tree diagram shown in Fig. 2 was obtained by the learning of the decision tree using 833 effective and 847 ineffective sequences. Then the prediction probabilities of gene silencing were computed for MG1 to MG5 by the learned decision tree diagram. The distributions of the predicted probabilities for MG1 to MG5 are shown in Fig. 3a–c. We also calculated the average for them.

The distributions of the predicted probabilities by decision tree learning, as a whole, indicated that there are differences between the effective and ineffective siRNA sequences as shown in Fig. 3a–c. There were clear distinctions between the distributions of effective and ineffective siRNAs for MG1 to MG5 except MG3. The entire average predicted probability of 103 effective siRNA sequences for these genes was 68.9%, whereas that of 93 ineffective siRNA sequences was 34.7%.

4.2. Evaluation of the Bayes' Theorem Method

The average probability predicted for these 833 siRNAs can be considered as a standard criterion for the effectiveness of siRNAs targeting other genes. The four cases listed in Table 5 were evaluated while computing the predicted gene silencing probabilities of

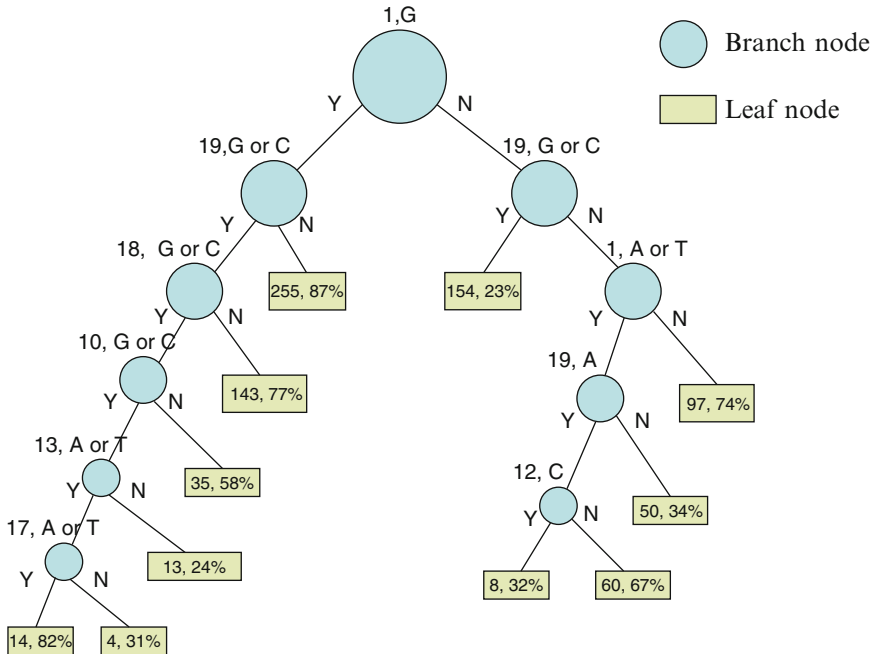


Fig. 2. Decision tree diagram for known 833 effective and 847 ineffective siRNA sequences. The *top* of the branch node indicates the position and nucleotide attribute, e.g., “19, G or C” means that the position of cDNA is 19 and the nucleotide at the position shows G or C. The *bottom* of the branch node shows yes (Y) and no (N). The leaf node indicates the number of effective siRNA sequences and its percentage, e.g., “255, 87%” means that the number of effective siRNA sequences is 255 and its percentage is 87% ($=255/292$).

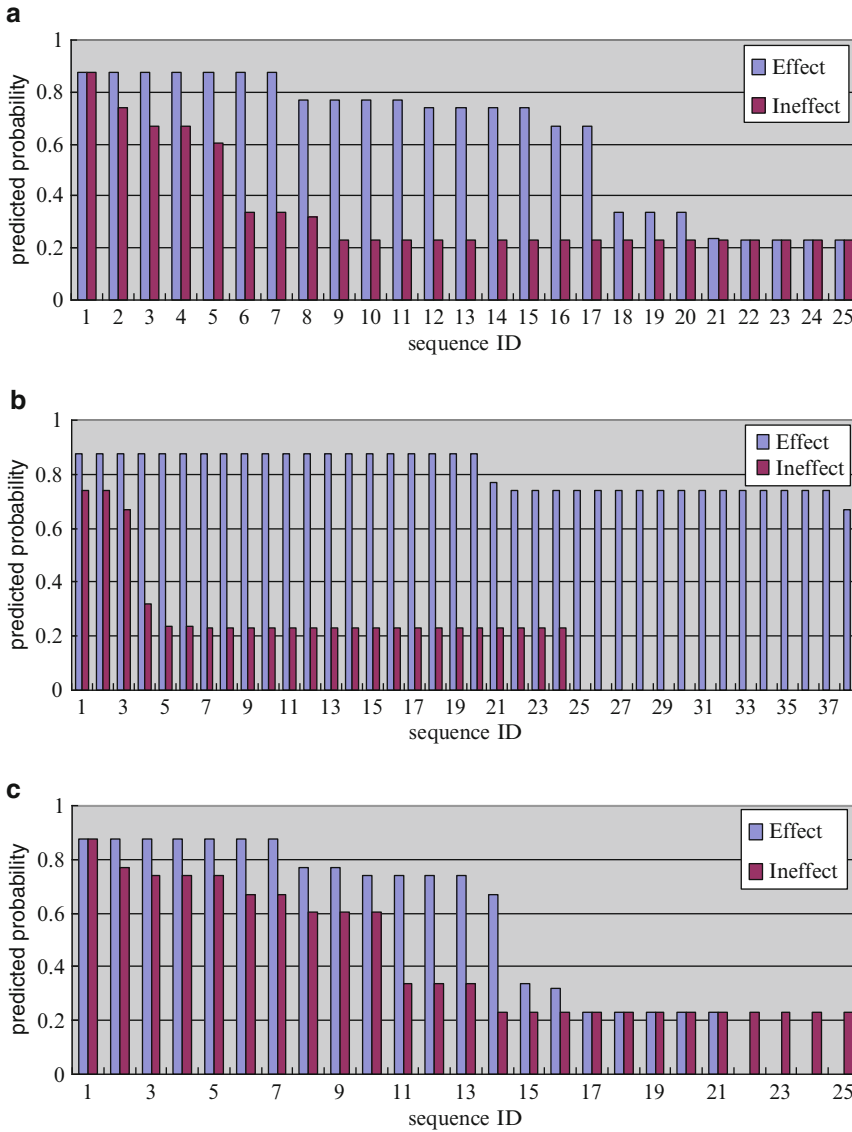


Fig. 3. Prediction probability distributions of siRNA sequences effective and ineffective for MG1 to MG5 by the proposed decision tree method. Effect: Effective siRNAs. Ineffect: Ineffective siRNAs. The probabilities for effective and ineffective siRNAs are computed by using the proposed decision tree method, and the siRNAs are sorted according to the predicted probabilities. They are numbered as sequence IDs (MG1: 1–25, MG2: 1–38, MG3: 1–25, MG4: 1–7, MG5: 1–12). (a) MG1 prediction distribution. (b) MG2 prediction distribution. (c) MG3 prediction distribution. (d) MG4 prediction distribution. (e) MG5 prediction distribution.

the effective (functional) and ineffective (nonfunctional) siRNAs for the recently reported genes (MG1 to MG5) under the prior probabilities of $P^{\text{eff}}=0.1$ and 0.2.

Because the probabilities predicted for individual siRNA sequences for the reported genes varied, the average predicted probability was used as an evaluation measure in the following verification. The reasonability of this is discussed later.

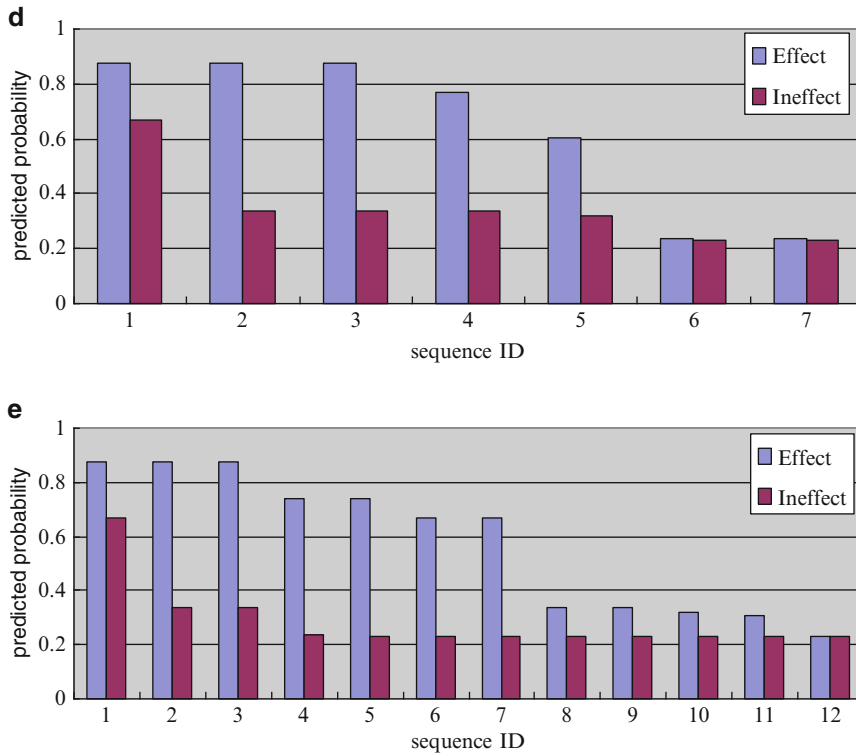


Fig. 3. (continued)

4.2.1. Case 1: Combination of Both Independent Nucleotide Occurrences

In Case 1, the average probability predicted that 833 effective siRNAs would be effective in silencing their target genes (Ae-833) was computed to be 29.9% by using Eq. 8 under the prior probability $P^{\text{eff}}=0.1$ (see Subheading 3.2). Although this probability might seem low for effective siRNAs, Bayes' theorem estimated it to be 2.99 times the prior probability. Therefore, this could be considered a standard criterion for the effectiveness of siRNAs. The distributions of the probabilities predicted for the effective and ineffective siRNA sequences for MG1 to MG5 are shown in Fig. 4. In the case of MG1, as shown in Fig. 4a, it is clear that there are big differences between the distributions for effective and ineffective siRNA sequences. Most of the effective siRNAs (88%), for example, have predicted probabilities of gene silencing that are >10%, whereas siRNAs more than 70% ineffective have predicted probabilities of effective gene silencing that are <10%. The average probability of effective silencing predicted for the effective siRNAs for MG1 (Ae-MG1) was 32.2%, whereas that predicted for the ineffective ones (Ai-MG1) was 6.7%. This means, according to Eq. 9, that the PR for sequences effective for MG1 is 3.22. That is, Bayes' theorem estimated a probability more than three times as high as the prior probability. The PR for Ai-MG1, in contrast, is 0.67, which indicates that the silencing obtained with sequences

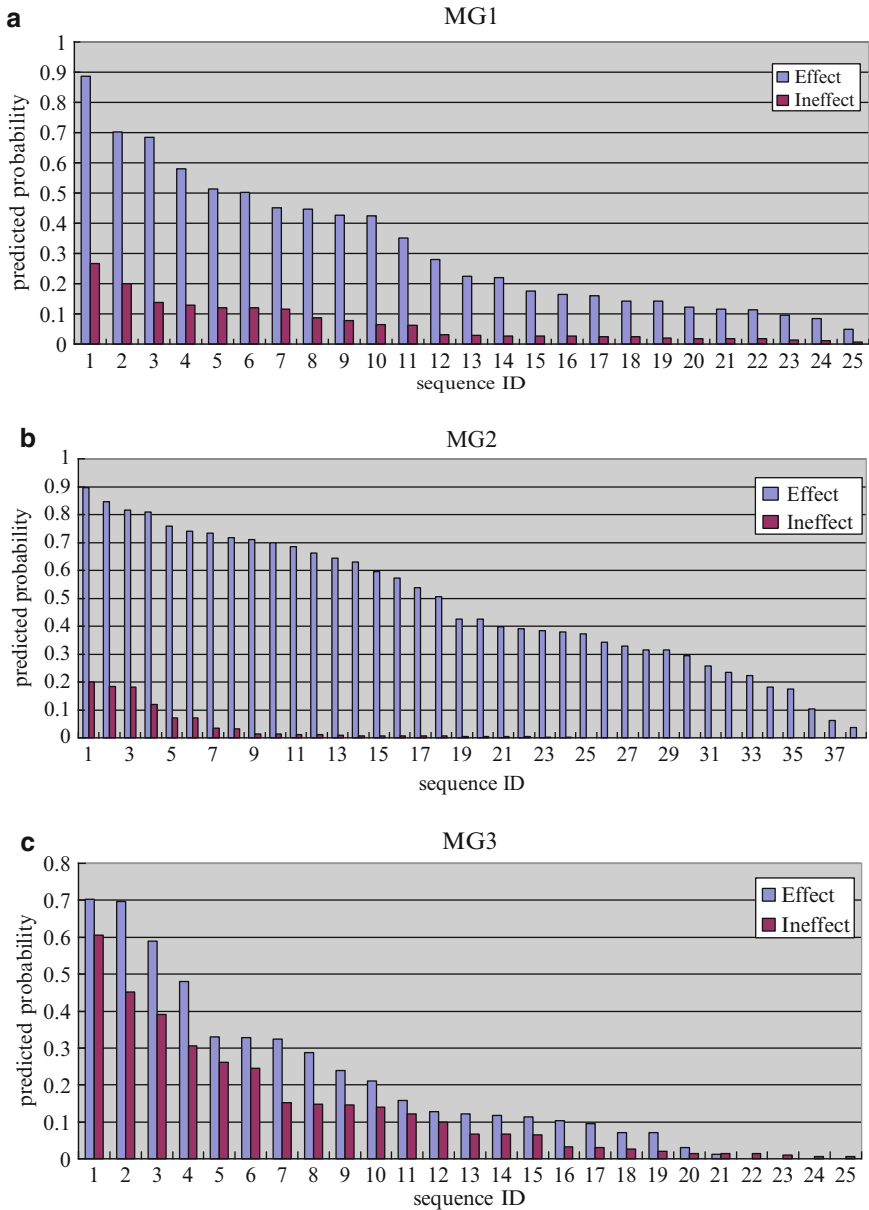


Fig. 4. Distributions of probabilities predicted for siRNA sequences that are effective and ineffective for MG1 to MG5 (results of Case 1 evaluation). Effect: Effective siRNAs. Ineffect: Ineffective siRNAs. The probabilities for effective and ineffective siRNAs were calculated by using Eq. 8, and the siRNAs are sorted according to the predicted probabilities. They are numbered as sequence IDs (MG1: 1–25, MG2: 1–38, MG3: 1–25, MG4: 1–7, MG5: 1–12). (a) Distribution predicted for MG1. (b) Distribution predicted for MG2. (c) Distribution predicted for MG3. (d) Distribution predicted for MG4. (e) Distribution predicted for MG5.

ineffective for MG1 sequences would be about two-third the usual level of silencing.

MG2 has remarkably distinct distributions. As shown in Fig. 4b, the gene silencing probabilities predicted for most siRNAs effectively silencing MG2 are high, whereas those predicted for

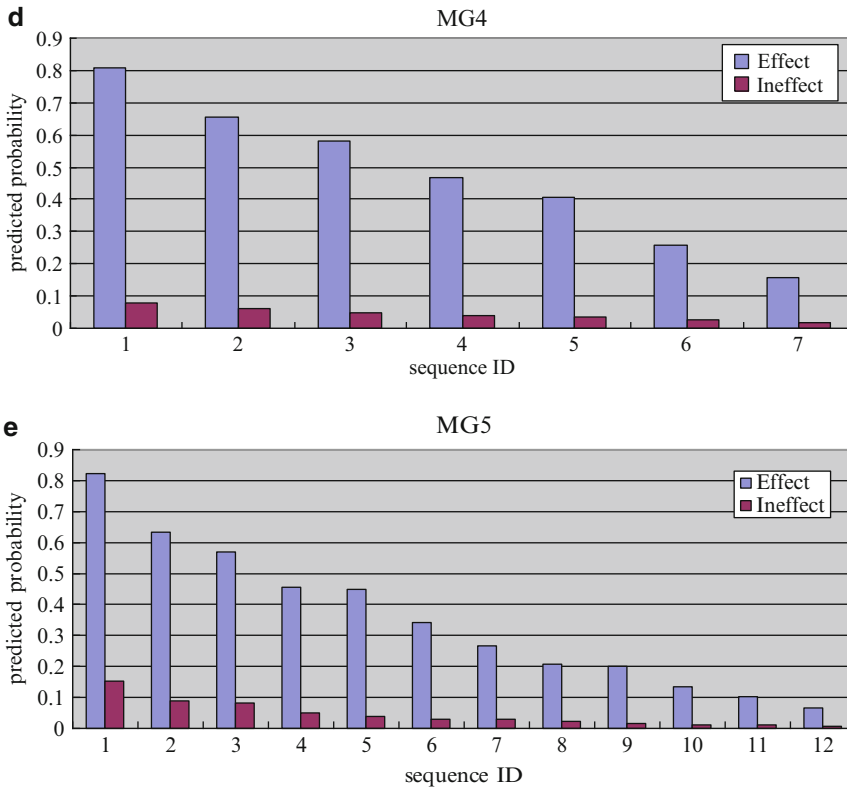


Fig. 4. (continued)

most of the ineffective ones are low. The average probability that predicted that the effective siRNAs for MG2 (Ae-MG2) would be effective was 47.9%, whereas the average probability that predicted that the ineffective ones would be effective (Ai-MG2) was 4.2%. These effective and ineffective estimation ratios are therefore better than those of MG1.

On the other hand, as shown in Fig. 4c, although the gene silencing probabilities predicted for most of the siRNAs effectively silencing MG3 are larger than those predicted for the ineffective ones, the differences are not so big as the differences between the probabilities predicted for the siRNAs effective and ineffective for MG1 and MG2. The average probability that predicted that the effective siRNAs for MG3 would be effective (Ae-MG3) was 24.8%, whereas the average probability that predicted that the ineffective ones would be effective (Ai-MG3) was 13.8%. This indicates that the *PRs* for siRNAs effective and ineffective for silencing MG3 (Ae-MG3 and Ai-MG3) are 2.48 and 1.38, so there might not be much difference between them. The average probabilities Ae-MG3 and Ai-MG3 reflect distribution features shown in Fig. 4c.

As shown in Fig. 4d, in contrast, the siRNAs effective and ineffective in silencing MG4 have remarkably distinct distributions.

The average probability that predicted that the effective siRNAs for MG4 would be effective (Ae-MG4) was 47.6%, whereas the average probability that predicted that ineffective ones would be effective (Ai-MG4) was 4.4%. This therefore indicates that the estimation ratios of the effective and ineffective siRNAs for MG4 are similar to those for MG2. The average probabilities of effective and ineffective gene silencing predicted for these siRNAs also reflect distribution features shown in Fig. 4d.

MG5 also has distinct distributions reflecting the average probabilities predicted for effective and ineffective gene silencing: Ae-MG5 = 35.4% and Ai-MG5 = 4.5% (Fig. 4e). There is therefore a better relation in the functional and nonfunctional ratios of MG1, MG2, MG4, and MG5. The entire average probability that predicted that the effective siRNAs for MG1 to MG5 would be effective was 37.9%, whereas the entire average probability that predicted that the ineffective ones would be effective was 7.7%.

With the prior probability $P^{\text{eff}} = 0.2$ the average probability that the 833 effective siRNAs would be effective was calculated to be 43.5%, 1.45 times larger than that calculated with $P^{\text{eff}} = 0.1$. The values listed in Table 6 show that with $P^{\text{eff}} = 0.2$ the average probability that siRNAs effective for MG1 to MG5 would be effective ranges from 1.32 to 1.46 times larger than that with $P^{\text{eff}} = 0.1$. On the other hand, with $P^{\text{eff}} = 0.2$ the average probability that the ineffective ones would be effective ranges from 1.67 to 2.1 times larger than that with $P^{\text{eff}} = 0.1$.

4.2.2. Case 2: Combination of Independent and Deductive Nucleotide Occurrences

In Case 2, Ae-833 was computed to be 25.1% under $P^{\text{eff}} = 0.1$. Ae-MG1 was 27.9%, whereas Ai-MG1 was 5.8%. These predictions mean that PRs of Ae-MG1 and Ai-MG1 are, respectively, 2.79 and 0.58. Ae-MG2 was 26.7%, whereas Ai-MG2 was 5.7%. As the PRs of Ae-MG2 and Ai-MG2 are, respectively, 2.67 and 0.57, they indicate levels similar to those of MG1. Because Ae-MG3 and Ai-MG3 were, respectively, 20.2% and 11.6%, their PRs are, respectively, 2.02 and 1.16. As Ae-MG4 and Ai-MG4, in contrast, were, respectively, 51.1% and 2.3%, their PRs are, respectively, 5.11 and 0.23. These predicted probabilities therefore indicate remarkably clearer distinctions than the prior probability. Because Ae-MG5 and Ai-MG5 were, respectively, 17.9% and 5.4%, their PRs are, respectively, 1.79 and 0.54. The entire average probability that predicted that the effective siRNAs for MG1 to MG5 would be effective was 26.3%, whereas the entire average probability that predicted that the ineffective ones would be effective was 7.1%. As a whole, the prediction accuracy of Case 2 is lower than that of Case 1. This indicates that the deductive nucleotide occurrences would not contribute to the prediction accuracy.

In the case of $P^{\text{eff}} = 0.2$, the average probability that 833 effective siRNAs would be effective was computed to be 38.5%, 1.54 times larger than that computed with $P^{\text{eff}} = 0.1$. The values listed in

Table 6
Average probabilities of effective silencing predicted for siRNAs effective and ineffective for various genes

Genes	Eff/Ineff	No. of seqs	Case 1		Case 2		Case 3		Case 4	
			$P^{\text{eff}} = 0.1$	$P^{\text{eff}} = 0.2$	$P^{\text{eff}} = 0.1$	$P^{\text{eff}} = 0.2$	$P^{\text{eff}} = 0.1$	$P^{\text{eff}} = 0.2$	$P^{\text{eff}} = 0.1$	$P^{\text{eff}} = 0.2$
MG 1	Effect	25	32.2	47	27.9	41.9	36	49.7	32.7	46.3
	Ineffect	25	6.7	13.1	5.8	11.6	9.5	15.5	7.7	13.7
MG 2	Effect	38	47.9	63.3	26.7	41.6	41	53.9	25	36.4
	Ineffect	24	4.2	8.2	5.7	10.4	1.4	2.9	2	4.2
MG 3	Effect	21	24.8	38	20.2	32.2	32.7	45.5	27.9	40.1
	Ineffect	25	13.8	22.9	11.7	20.8	15	23.1	13.9	22
MG 4	Effect	7	47.6	64	51.1	68.9	65.1	77.7	70.2	82.7
	Ineffect	7	4.4	9.2	2.3	5	5.9	12	3.1	6.7
MG 5	Effect	12	35.4	50.7	17.9	30	36.1	49.8	19.8	31.4
	Ineffect	12	4.5	9.3	5.5	11	5.1	9.1	5.9	10.4
siRNAs	Effect	833	29.9	43.5	25.1	38.5	38.4	51.6	33.7	46.9

Effect: Effective siRNAs, ineffect: ineffective siRNAs, P^{eff} : prior probability
No. of seqs: Number of siRNA sequences

Table 6 show that the average probability that effective siRNAs for MG1 to MG5 would be effective ranges from 1.35 to 1.68 times larger than that in case of $P^{\text{eff}}=0.1$. On the other hand, the average probability that the ineffective ones would be effective ranges from 1.8 to 2.2 times larger than that in case of $P^{\text{eff}}=0.1$.

4.2.3. Case 3: Combination of Markov Model and Independent Nucleotide Occurrences

In Case 3, Ae-833 was computed to be 38.4% under the assumption that $P^{\text{eff}}=0.1$. This is a higher probability than the corresponding Ae-833 computed in Case 1. Ae-MG1 was 36%, whereas Ai-MG1 was 9.5%. As the PRs of Ae-MG1 and Ai-MG1 are, respectively, 3.6 and 0.95, Ae-MG1 indicates better accuracy than that in Case 1 and Ai-MG1 shows worse accuracy than that in Case 1. Ae-MG2 was 40.95%, whereas Ai-MG2 was 1.37%. So the MG2 of Case 3 indicates that Ae-MG2 has a worse accuracy than that in Case 1, whereas Ai-MG2 has an accuracy three times better than that in Case 1. Because Ae-MG3 and Ai-MG3 were, respectively, 32.7% and 15%, their PRs are, respectively, 3.27 and 1.5. So the MG3 of Case 3 indicates that Ae-MG3 is better than that in Case 1, whereas Ai-MG3 is little bit worse than that in Case 1. As Ae-MG4 and Ai-MG4 in Case 3 were, respectively, 65.1% and 5.86%, Ae-MG4 is better than that in Case 1 and Ai-MG4 is a little bit worse than that in Case 1. Because Ae-MG5 and Ai-MG5 are, respectively, 36.1% and 5.1%, their PRs are similar to those in Case 1. As a whole, the PRs in Case 3 indicate better accuracy than do those in Case 1.

With $P^{\text{eff}}=0.2$ the average probability that 833 effective siRNAs would be effective was computed to be 51.6%, 1.34 times larger than that computed with $P^{\text{eff}}=0.1$. The values listed in Table 6 show that with $P^{\text{eff}}=0.2$ the average probability that siRNAs effective for MG1 to MG5 would be effective ranges from 1.19 to 1.39 times larger than that with $P^{\text{eff}}=0.1$. On the other hand, with $P^{\text{eff}}=0.2$ the average probability that the ineffective ones would be effective ranges from 1.5 to 2.1 times larger than that with $P^{\text{eff}}=0.1$.

4.2.4. Case 4: Combination of Markov Model and Deductive Nucleotide Occurrences

In Case 4, Ae-833 was computed to be 33.7% under $P^{\text{eff}}=0.1$. This is 4.7% lower than the corresponding Ae-833 computed in Case 3. Ae-MG1 was 32.7%, whereas Ai-MG1 was 7.7%. The PRs of Ae-MG1 and Ai-MG1 are therefore, respectively, 3.27 and 0.77. Ae-MG2 was 25%, whereas Ai-MG2 was 2%. Because Ae-MG3 and Ai-MG3 were, respectively, 27.9% and 13.9%, their PRs are 2.79 and 1.39. As Ae-MG4 and Ai-MG4 in Case 4 were, respectively, 70.2% and 3.1%, they are distinguished more clearly in this case than in any other. Because Ae-MG5 and Ai-MG5 were, respectively, 19.8% and 5.94%, their PRs are 1.98 and 0.594. As a whole, although the evaluation results of MG1 to MG5 in Case 4 showed better accuracy than those in Case 2, they showed worse accuracy than those in Cases 1 and 3.

In the case of $P^{\text{eff}}=0.2$, the average probability that 833 effective siRNAs would be effective was computed to be 46.9%, 1.39 times larger than that computed with $P^{\text{eff}}=0.1$. The values listed in

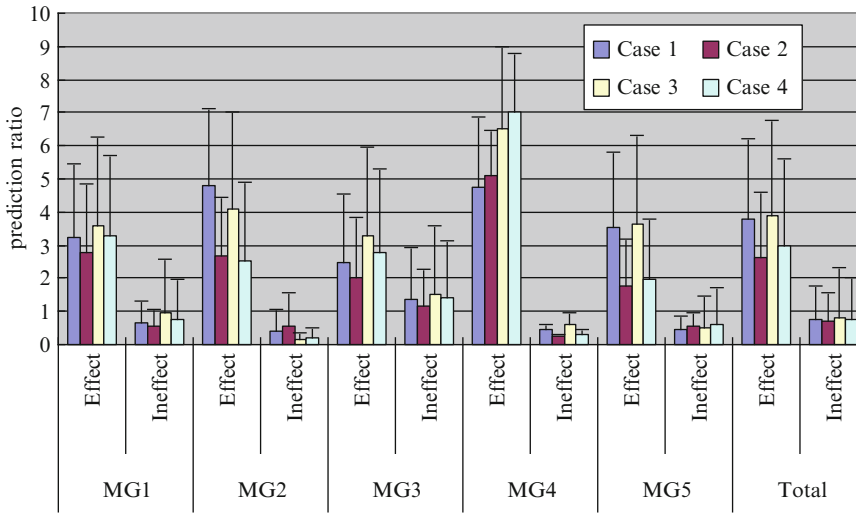


Fig. 5. Relations between the predicted probability and the prior probability. Effect: Effective siRNAs. Ineffect: Ineffective siRNAs. The prediction ratio graphed here is the ratio of the average of the probabilities estimated for the siRNAs to the prior probability $P^{\text{eff}} = 0.1$. Case 1 is that in which both the effective and ineffective siRNAs have independent nucleotide occurrences at the individual positions. Case 2 is that in which the effective siRNAs have independent and ineffective siRNAs have deductive nucleotide occurrences at the individual positions. Case 3 is that in which the effective siRNAs have dependent nucleotide occurrences at the individual positions (the simple Markov model) and the ineffective siRNAs have independent nucleotide occurrences at the individual positions. Case 4 is that in which the effective siRNAs have dependent nucleotide occurrences and the ineffective siRNAs have deductive nucleotide occurrences at the individual positions.

Table 6 show that the average probability that siRNAs effective for MG1 to MG5 would be effective ranges from 1.18 to 1.59 times larger than that in case of $P^{\text{eff}} = 0.1$. On the other hand, the average probability that the ineffective ones would be effective ranges from 1.6 to 2.1 times larger than that in case of $P^{\text{eff}} = 0.1$.

The ratios of the average estimated probabilities to the prior probabilities of the reported genes are shown for Cases 1–4 in Fig. 5, where it is clear that the ratios computed in Case 3 show the most distinct differences between the siRNAs effective and ineffective for silencing MG1 to MG5 except MG2. This means that the combination of Markov model and independent nucleotide occurrences might yield the most accurate predictions when Bayes' theorem is used to select siRNA sequences effective for gene silencing. With regard to gene classes, MG1 and MG5 show distinctions between the effective and ineffective siRNAs more clearly than MG3 does, and MG2 and MG4 show distinctions remarkably clearly. These results therefore imply that there are some differences in the individual nucleotide frequencies at each position of the siRNAs effective for these gene classes. Although MG3 shows differences between the effective and ineffective siRNAs, the ratios of the predicted probabilities for effective ones to ineffective ones are < 2.2 . This implies that there is no big difference between the individual nucleotide frequencies of the siRNAs effective and ineffective for silencing this class of genes.

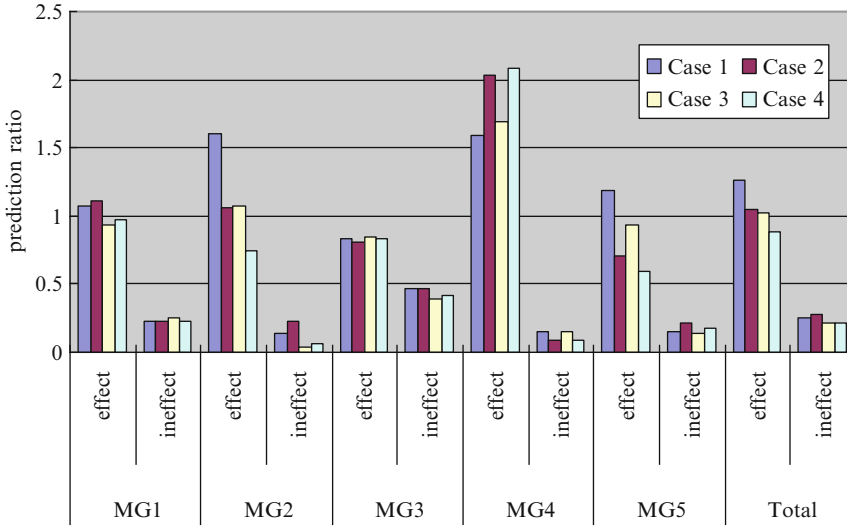


Fig. 6. Normalized relations among the siRNAs for MG1 to MG5. Effect: Effective siRNAs. Ineffect: Ineffective siRNAs. The prediction ratio graphed here is the ratio of the average estimated for the designated siRNAs to the average for the 833 siRNAs known to be effective. Case 1 is one in which both the effective and ineffective siRNAs have independent nucleotide occurrences at the individual positions under the assumption that $P^{\text{eff}} = 0.1$ when the 833 siRNAs known to be effective are used as normalization data. Case 2 is also one in which P^{eff} is assumed to be 0.1 when the 833 siRNAs known to be effective are used as normalization data, but in this case the effective siRNAs have independent nucleotide occurrences and ineffective siRNAs have deductive nucleotide occurrences. Case 3 is one in which the effective siRNAs have dependent (simple Markov model) nucleotide occurrences and the ineffective siRNAs have independent nucleotide occurrences at the individual positions. Case 4 is one in which the effective siRNAs have dependent nucleotide occurrences and the ineffective siRNAs have deductive nucleotide occurrences at the individual positions.

4.2.5. Comparative Analysis by the Normalization

Because the average predicted probability that 833 effective siRNAs would be effective (Ae-833) could be considered a standard criterion for the other gene functionality as described earlier, we calculated NRs for MG1 to MG5 by using Eq. 10 (see Subheading 3.2). As a whole, individual NRs result in clearer distinctions between the effective and ineffective siRNAs. As shown in Fig. 6, in all Cases 1–4 the average NRs of siRNAs effective for MG1 are about 1. These results therefore indicate that all 833 siRNAs effective for MG1 are about equally effective. Because the average NRs of the 833 siRNAs effective for MG2 that are calculated in three Cases 1–3 are more than 1.1, they indicate that these siRNAs are more likely to be effective than the 833 siRNAs known to be effective. In contrast, the average NRs of siRNAs effective for MG3 are about 0.83 in all cases, whereas those of the siRNAs ineffective for MG3 are in the ranges of 0.39–0.47. Therefore, although there are differences between the average NRs for siRNAs effective and ineffective for MG3, the ratios of between them are 1.73–2.18 and do not indicate differences as clearly as do the average NRs for siRNAs effective and ineffective for MG1 and MG2. Meanwhile, in all cases the average NRs of siRNAs effective for MG4 are about 1.6 times higher than the Ae-833 standard criterion. These results imply that

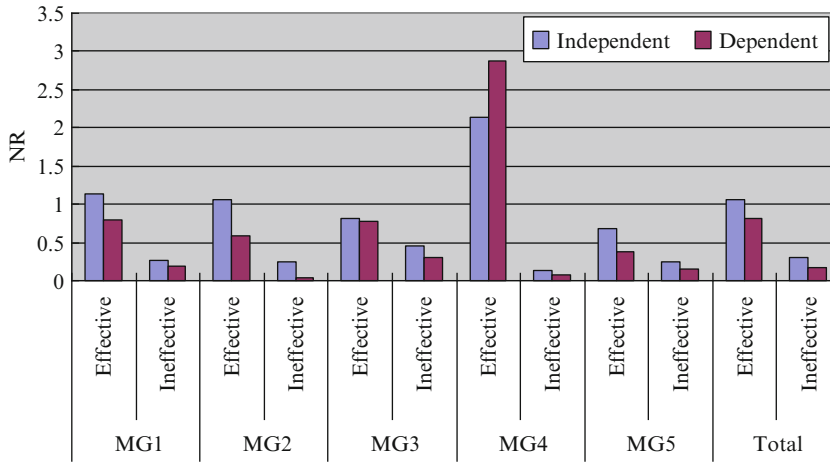


Fig. 7. Normalized ratios based on 833 effective siRNAs. Effective: Effective siRNAs. Ineffective: Ineffective siRNAs. NR: Normalized ratio calculated by Eq. 22. Independent: Independent occurrences at individual positions. Dependent: The simple Markov model.

the siRNAs effective for MG4 are potentially more effective than the 833 siRNAs known to be effective. The average NRs of siRNAs effective for MG5 range from 0.59 to 1.18, whereas those of ineffective siRNAs for MG5 are 0.13–0.22. The average NRs of all the siRNAs effective for MG1 to MG5 are more than 1 in three Cases 1, 2, and 3, indicating that these siRNAs are potentially more effective than the 833 siRNAs known to be effective. On the other hand, the average NRs of the siRNAs ineffective for MG1 to MG5 are in the ranges of 0.21–0.28. They therefore contribute to the clearer differences between the average NRs of the siRNAs effective and ineffective for MG1, MG2, MG4, and MG5 (Fig. 6). This makes the selection of candidate effective siRNA sequences easier.

4.3. Evaluation of the Average Silencing Probability Method

The proposed average silencing probability method was evaluated by first computing A_E for 833 effective siRNA sequences and then using Eq. 14 to compute the individual probabilities of the effective and ineffective siRNAs. Because 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 Fig. 7.

4.3.1. 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.

Because 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). Because 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 shows quite low likelihood (0.13 times). On the other hand, because the normalized ratios for MG5 were, respectively, 0.68 and 0.24, the ratio of the effective to 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 Fig. 7, 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

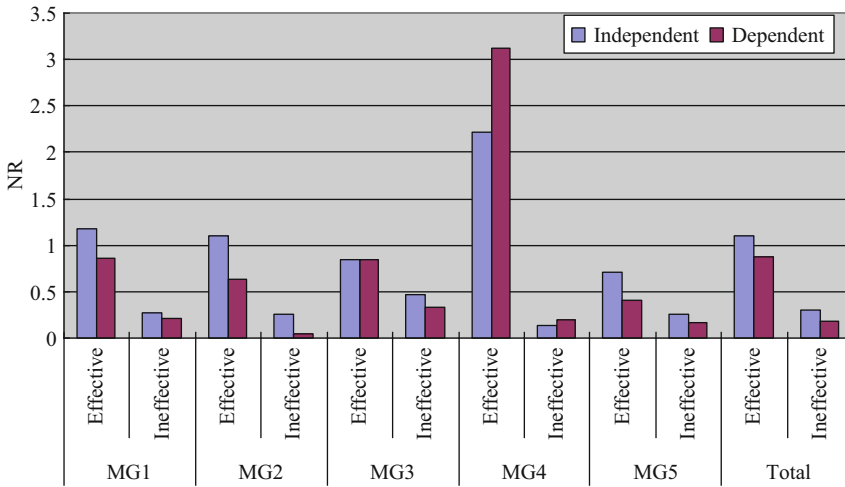


Fig. 8. Normalized ratios based on 636 effective siRNAs.

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, because the average normalized ratios of MG4 (cyclin B1) effective and ineffective siRNAs were, respectively, 2.88 and 0.08, the difference between them was 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.

4.3.2. 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 these siRNAs and the probabilities that the effective and ineffective siRNAs would be effective for the reported genes are shown in Fig. 8. Although there were ups and downs in *NR*s predicted for MG1 to MG5 using either the 833 or 636 effective siRNAs, the total *NR*s predicted are similar for both cases. That is, the *NR*s 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.

4.4. Evaluation of the HMM Method

The Viterbi algorithm was carried out for the state diagram of the HMM shown in Fig. 1. As a result, the siRNA sequence GAAGA

AGAGAGAGAGCAGA was obtained as the optimal nucleotide sequence (i.e., the sequence maximizing the sequence state probability for positions 1–19). 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 Eqs. 16–21, and yielded the sequence TTTTATT AATCGCGTTCG. 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 7 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 13 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 11 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.

4.4.1. 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 Subheading 3.5). The probabilities of individual nucleotide occurrence frequencies at individual positions are listed in Table 4c. The relations among NRs of effective and ineffective siRNAs for MG1 to MG5 computed by using the Eqs. 14, 15, and 22 are shown in Fig. 9. In the case of using the 847 known ineffective siRNAs, NRs of effective siRNAs for MG1 to MG5 are <1, whereas those of ineffective ones are more than 1 as shown in Fig. 9. The NR of the total effective siRNAs is 0.67, whereas that of the ineffective ones is 2.37.

Comparing Fig. 9 with Fig. 7, it is clear that the corresponding NRs of effective and ineffective siRNAs for MG1 to MG5 are, respectively, reverse relations. This depends on what set of siRNAs,

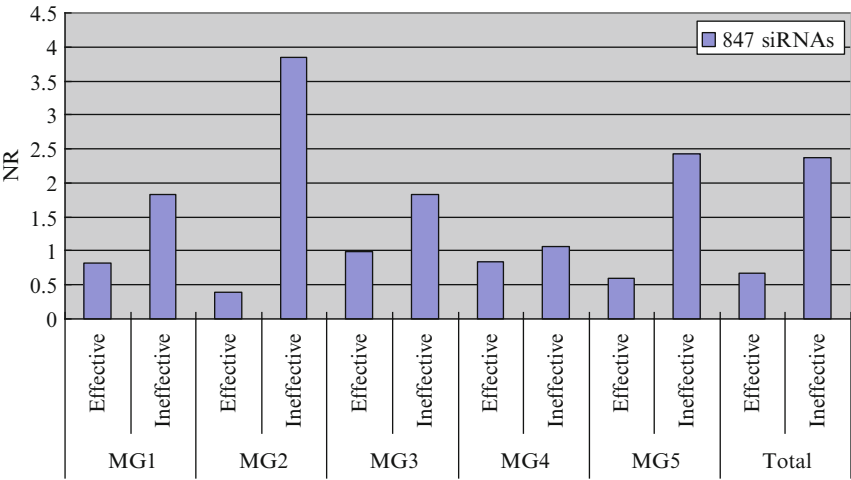


Fig. 9. Normalized ratios based on 847 ineffective siRNAs.

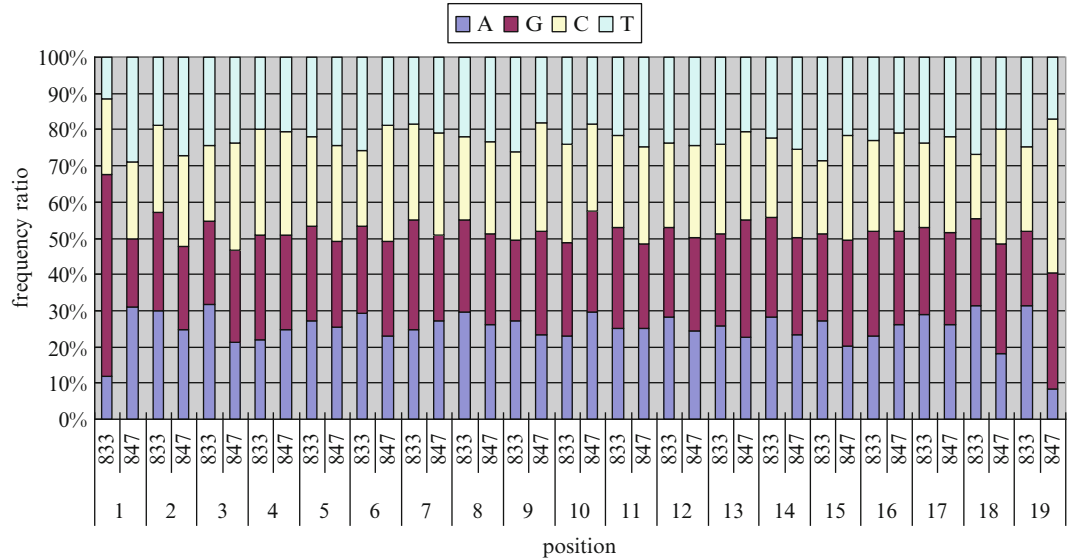


Fig. 10. Relations of nucleotide occurrence frequencies between 833 effective and 847 ineffective siRNAs.

i.e., 833 or 847 siRNAs, is used. There are differences in the nucleotide occurrence frequencies between both sets of siRNAs as shown in Fig. 10. Especially, there are big differences at positions 1 and 19. These results are also useful for designing effective siRNA sequences.

4.4.2. Characteristics for the Combinations of Two Successive Nucleotides

From the relations between two successive nucleotides determined by using the first Markov model for 833 known effective siRNAs it is possible to analyze the frequencies of combinations of two successive nucleotides in the sense strand. The relations among the

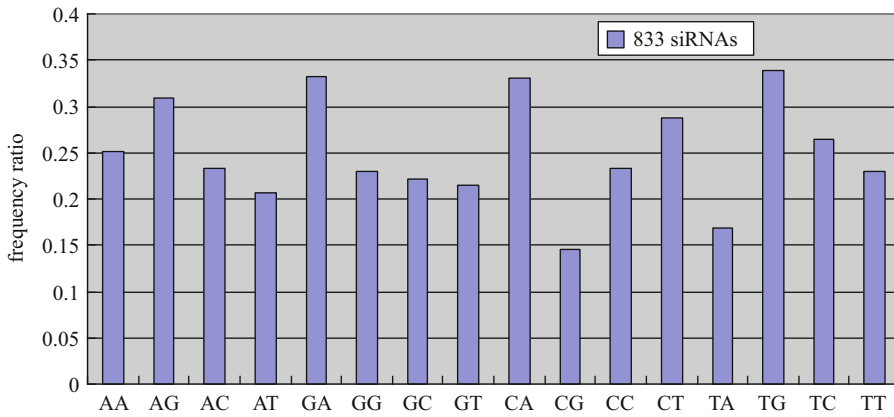


Fig. 11. Frequency ratios of combinations of two successive nucleotides in 833 effective siRNAs.

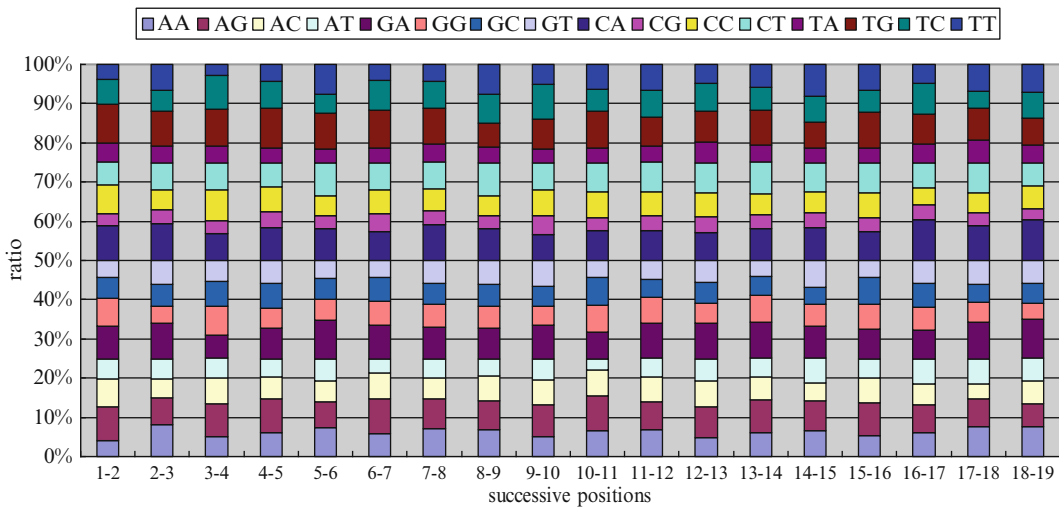


Fig. 12. Frequency ratios of two-nucleotide combinations in two successive positions from 5' to 3' for 833 effective siRNAs.

frequency ratios of two successive nucleotides for 833 known effective siRNAs are shown in Fig. 11, where it is clear that there are ups and downs in the frequency ratios of combinations of two nucleotides. Two-nucleotide combinations with high frequency ratios are TG (34%), GA (33%), CA (33%), and AG (31%), whereas combinations with low ones are CG (15%) and TA (17%).

It is also possible to calculate the frequency ratios of two-nucleotide combinations between two successive positions from 5' to 3' of the sense strand. They are shown in Fig. 12. When designing effective siRNAs for the target genes, it is also necessary to consider these characteristics of the frequencies of two-nucleotide combinations.

References

1. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (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, Tuschl T (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in mammalian cell culture. *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, Navia CD, Sharp PA (2003) Killing the messenger: short RNAs that silence gene expression. *Nat Rev* 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, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD (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, Sonnhhammer ELL (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–W129
16. Santoyo J, Vaguerizas JM, Dapozo J (2004) Highly specific and accurate selection of siRNAs for high-throughput functional assays. *Bioinformatics* 21:1376–1382
17. Truss M, Swat M, Kielbasa SM, Schafer R, Herzed H, Hagemeier C (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, Khvorova A (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, Juni A, Ueda R, Saigou K (2004) Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res* 32: 936–948
20. Amarzguioui M, Prydz H (2004) An algorithm for selection of functional siRNA sequences. *Biochem Biophys Res Commun* 316: 1050–1058
21. Hsieh AC, Bo R, Monola J, Vazquez F, Bare O, Khvorova A, Scaringe S, Sellers WR (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, Zatorski A, Shum D, Mayer T, De Angelis DA, Ouerfelli O, Rutishauser U, Rothman JE (2005) Sequence characteristics of functional siRNAs. *RNA* 11:864–872
23. Huesken D, Lange J, Mikanin C, Weiler J, Asselbergs F, Warner J, Meloon B, Engel S, Rosenberg A, Cohen D, Labow M, Reinhardt M, Natt F, Hall J (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, Grunfeld T, Saetrom PO (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, Cambridge
26. Takasaki S (2009) Selecting effective siRNA target sequences by using Bayes' theorem. *Comput Biol Chem* 33:368–372
27. Takasaki S, Kawamura Y, Konagaya A (2006) Selecting effective siRNA sequences by using radial basis function network and decision tree learning. *BMC Bioinform* 7(Suppl 5):S22

28. Takasaki S, Kotani S, Konagaya A (2004) An effective method for selecting siRNA target sequences in mammalian cells. *Cell Cycle* 3: 790–795
29. Takasaki S, Kotani S, Konagaya A (2005) Selecting effective siRNA target sequences for mammalian genes. *RNA Biol* 2:21–27
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-Bioinform J* 6:69–84
32. Takasaki S, Kawamura Y (2007) Using radial basis function networks and significance testing to select effective siRNA sequences, *Comput. Stat Data Anal* 51:6476–6487
33. 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
34. Ladunga I (2007) More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. *Nucleic Acids Res* 35:433–440
35. Holen T (2006) Efficient prediction of siRNAs with siRNA rules 1.0: an open-source JAVA approach to siRNA algorithms. *RNA* 12: 1620–1625
36. Heale BSE, Sifer HS, Bowers C, Rossi JJ (2005) siRNA target site secondary structure predictions using local stable substructures. *Nucleic Acids Res* 33:e-30
37. Luo KQ, Chang DC (2004) The gene silencing efficacy of siRNA is strongly dependent on the local structure of mRNA at the target region. *Biochem Biophys Res Commun* 318:303–310
38. Bohula EA, Salisbury AJ, Sohail M, Playford MP, Riedemann J, Southern EM, Macaulay VM (2003) The efficacy of small interfering RNAs targeted to the type I insulin-like growth factor receptor (IGFIR) is influenced by secondary structure in the IGFIR transcript. *J Biol Chem* 278:15991–15997
39. Chan CY, Carmack CS, Long DD, Maliyekkel A, Shao Y, Roninson IB, Ding Y (2009) A structural interpretation of the effect of GC-content on efficiency of RNA interference. *BMC Bioinform* 10(Suppl 1):S33
40. Vig K, Lewis N, Moore EG, Pillai S, Dennis VA, Singh SR (2009) Secondary RNA structure and its role in RNA interference to silence the respiratory syncytial virus fusion protein gene. *Mol Biotechnol* 43:200–211
41. Saetrom P, Snove O Jr (2004) A comparison of siRNA efficacy predictors. *Biochem Biophys Res Commun* 321:247–253
42. Shabalina SA, Spiridonov AN, Ogurtsov AY (2006) Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinform* 7:65
43. Vert J, Foveau N, Lajaunie C, Vandenbrouck Y (2006) An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinform* 7:520
44. Matveeva O, Nechipurenko Y, Rossi L, Moore B, Saetrom P, Ogurtsov AY, Atkins JF, Shabalina SA (2007) Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. *Nucleic Acids Res* 35:e63
45. Lu ZJ, Mathews DH (2008) Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res* 36:640–647
46. Wang X, Wang X, Varma RK, Beauchamp L, Magdaleno S, Sendera TJ (2009) Selection of hyperfunctional siRNAs with improved potency and specificity. *Nucleic Acids Res* 37:e152
47. Klingelhoefer JW, Moutsianas L, Holmes C (2009) Approximate Bayesian feature selection on a large meta-dataset offers novel insights on factors that effect siRNA potency. *Bioinformatics* 25:1594–1601
48. Gong W, Ren Y, Zhou H, Wang Y, Kang S, Li T (2008) siDRM: an effective and generally applicable online siRNA design tool. *Bioinformatics* 24:2405–2406
49. Patzel V (2007) In silico selection of active siRNA. *Drug Discov Today* 12:139–148
50. Tafer H, Ameres SL, Obemosterer G, Gebeshuber CA, Schroeder R (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat Biotechnol* 26: 578–583
51. Walton SP, Wu M, Gredell JA, Chan C (2010) Designing highly active siRNAs for therapeutic applications. *FEBS J* 277:4806–4813
52. Ahmed F, Raghava GP (2011) Designing of highly effective complementary and mismatch siRNAs for silencing a gene. *PLoS One* 6: e23443
53. Chaudhary A, Srivastava S, Garg S (2011) Development of a software tool and criteria evaluation for efficient design of small interfering RNA. *Biochem Biophys Res Commun* 404: 313–320
54. Katoh T, Suzuki T (2007) Specific residues at every third position of siRNA shape its efficient RNAi activity. *Nucleic Acids Res* 35:e27
55. Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1:81–106
56. Saetrom P (2004) Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinformatics* 20:3055–3063



<http://www.springer.com/978-1-62703-118-9>

siRNA Design

Methods and Protocols

Taxman, D.J. (Ed.)

2013, XIII, 392 p., Hardcover

ISBN: 978-1-62703-118-9

A product of Humana Press