Chem-Bio Informatics Journal, Vol. 6, No. 3, pp.69-84(2006)

Comparative analyses for selecting effective siRNA sequences

Shigeru Takasaki[1,*] and Akihiko Konagaya[1]

[1] *RIKEN Genomic Sciences Center (GSC), Suehiro-cho 1-7-22-E216, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan*
[*] *Correspondence should be addressed to Shigeru Takasaki*

*E-mail: *takasaki@gsc.riken.jp*

(Received November 17, 2005; accepted October 26, 2006; published online November 8, 2006)

Abstract

Because the short interfering RNA (siRNA) widely used for studying gene functions in mammalian cells varies markedly in its gene-silencing efficacy, several siRNA design rules/guidelines have been reported recently. Analyzing the reported siRNA design guidelines from qualitative and quantitative points of views, we found that they were not always effective selection rules for many other mammalian genes. Though some rules from the guidelines are suitable for extracting effective sequences for specific genes, they might sometimes be unsuitable for selecting sequences for other genes. Since the gene-silencing efficacy depends very much on the target sequence positions selected from the target gene, we examined 860 effective siRNA sequences from 503 different mammalian cDNAs in the literature. As a result, we got many preferred and unpreferred nucleotides different from the ones used in the previous guidelines. These sequence-dependent nucleotides could be used as a more general guideline for selecting new siRNA sequences in target genes. We proposed a measure (score) for selecting effective siRNA candidates based on the positional features of specific significant nucleotides and demonstrated the effectiveness of the proposed measure compared with the recently reported other selection methods. In this paper we also discussed the elimination of ineffective siRNA sequences from target candidates and optimal GC content in siRNA sequences.

Key Words: siRNA, RNA interference, siRNA guideline, gene-silencing, target sequence

Area of Interest: Bioinformatics and Bio Computing

Chem-Bio Informatics Journal, Vol. 6, No. 3, pp.69-84(2006)

1. Introduction

RNA interference (RNAi) silences gene expression by introducing double-stranded RNA (dsRNA) homologous to the target mRNA, and it has been widely used for studying gene functions [4] [5] [8] [19]. The short interfering RNA (siRNA) responsible for RNA interference, however, varies markedly in its gene-silencing efficacy in mammalian genes, where the gene-silencing effectiveness depends very much on the target sequence positions (sites) selected from the target gene [6] [10]. That is, different siRNAs synthesized for various positions induce different levels of gene-silencing. This indicates that the selection of the target sequence is critical to the effectiveness of the siRNA and that we therefore need useful criteria for gene-silencing efficacy when we are designing siRNA sequences. Some of the factors related to gene-silencing efficacy that have reported by investigators studying RNAi molecular mechanisms for mRNA cleavage are the binding energy, "GC" content, point-specific nucleotides and specific motif sequences [12] [18]. Since these factors could therefore be inferred to play important roles in determining effective siRNA sequences for the target gene, several siRNA design rules/guidelines using these factors have reported [1] [2] [11] [15] [24]. Although a comparative analysis of them has been reported recently [14] [16] [17] [20] [23], there is no consideration of the most important nucleotide features. Furthermore, although there is an off-target regulation risk in RNAi, when we use gene-silencing for studying gene functions we have to first somehow select high-potential siRNA candidates and eliminate possible off-target candidates.

We therefore examined three typical recently reported guidelines for siRNA efficacy [1] [15] [24] from the points of nucleotide occurrences for effective siRNAs and found few consistencies among them, possibly because they are based on sequence analyses for only a few target genes. Evaluating these guidelines qualitatively and quantitatively, we concluded that they are not always useful for selecting highly effective siRNA sequences for genes other than the target genes. Although there are other rules for siRNA designs, they seem to be not so clear factors for effective siRNAs [16] [17] [23]. What we need is a more general guideline for selecting the siRNA target sequence. Hypothesizing that the nucleotide occurrence trends are important, we examined previously reported effective siRNA sequences for clear tendencies in nucleotide occurrence. We examined 860 effective siRNA target sequences from 503 different mammalian cDNAs in the literature in the PubMed database [3] [9] [13]. Analyzing these sequences statistically, we found important features other than the ones used in the previous guidelines [1] [11] [15] [24].

This paper will first clarify the effectiveness of previous guidelines qualitatively and quantitatively. It will then describe positional features of specific significant nucleotides found by analyzing 860 sequences and will propose a measure (score) for selecting effective siRNA candidates based on the obtained positional features of significant nucleotides and will evaluate the effectiveness of the proposed measure compared with the other selection methods using scores [17]. It will finally discuss the elimination of ineffective sequences from target candidates and the optimal "GC" content of siRNA sequences.

2. Materials and Methods

2.1 Relations between individual guidelines and the effective/ineffective siRNA sequences

Individual guidelines are summarized in Figure 1. Guideline 1 (G1) specifies four preferred nucleotides: I (A at position 3), II (T at position 10), III (A or C or T at position 13) and IV (A or T at position 19) [15]. As the preferred nucleotides should occur at these positions with higher

probabilities than other nucleotides, we compared their occurrence probabilities there with their average occurrence probabilities in the effective and ineffective target sequence populations. We carried out similar comparisons for the following Guidelines 2 and 3. Guideline 2 (G2) specifies two preferred nucleotides, I (G or C at position 1) and II (A or T at position 19) and two unpreferred nucleotides, III (A or T at position 1) and IV (G or C at position 19) [24]. Guideline 3 (G3) specifies five preferred nucleotides — I (G or C at position 1), II (A at position 6), III (T at position 13), IV (C at position 16) and V (A or T at position 19) — and specifies three unpreferred nucleotides: VI (T at position 1), VII (T at position 10) and VIII (G at position 19) [1].

As G1, G2 and G3 are respectively based on the analyses of two genes (firefly luciferase and human cyclophilin B), six genes (firefly luciferase (PRL-TK), Vimentin, Oct 4, EGFP, ECFP, and DsRed) and four genes (human tissue factor (hTF), murine tissue factor (mTF), human protein serine kinase H1 (PSK) and human C-Src tyrosine kinase (CSK)), for simplicity these genes are symbolized as MG1-1 (firefly luciferase), MG1-2 (human cyclophilin B), MG2 (six genes) and MG3 (four genes) throughout this paper. Effective and ineffective siRNA sequences for the symbolized genes were selected from the literature in the following way [1] [15] [22] [24].

MG1-1: 25 effective and 25 ineffective sequences, MG1-2: 25 effective and 25 ineffective sequences, MG2: 38 effective and 24 ineffective sequences and MG3: 21 effective and 25 ineffective sequences.

To get a large number of effective siRNA sequences, we collected target sequences from published references in the PubMed database. As a result, we obtained 860 effective siRNA sequences (more than 70% gene-silencing) from 503 different cDNAs. The numbers of individual nucleotide occurrences at each of positions from 1 to 19 of MG1-1, MG1-2, MG2, MG3 and 503 gene effective sequences are respectively listed in Tables 1A, 1B, 1C, 1D and 1E.

Table 1. MG1-1, MG1-2, MG2, MG3 and 503 gene effective sequence distributions.

1A. MG1-1 Effective sequences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	4	6	7	6	8	2	10	6	8	7	9	7	10	4	11	11	11	8	12
G	10	8	9	6	7	9	6	10	6	6	5	13	4	14	9	6	5	5	5
C	3	5	3	7	3	6	4	6	6	7	6	3	5	4	3	4	5	4	3
T	8	6	6	6	7	8	5	3	5	5	5	2	6	3	2	4	4	8	5

1B. MG1-2 Effective sequences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	4	11	11	7	7	11	11	10	5	8	6	6	8	8	9	4	9	8	10
G	12	6	6	3	6	7	6	8	7	3	8	9	8	8	8	10	7	7	3
C	7	5	1	11	6	2	5	3	9	6	6	5	3	3	5	6	3	2	6
T	2	3	7	4	6	5	3	4	4	8	5	5	6	6	3	5	6	8	6

1C. MG2 Effective sequences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	1	10	9	9	13	12	16	9	11	8	8	7	17	16	16	18	15	17	23
G	21	4	10	10	7	7	11	8	12	5	10	16	1	4	4	4	6	4	0
C	16	15	9	11	11	12	8	12	8	14	10	10	2	6	6	1	6	2	1
T	0	9	10	8	7	7	3	9	7	11	10	5	18	12	12	15	11	15	14

1D. MG3 Effective sequences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	3	5	5	3	2	10	4	7	5	6	9	7	6	6	11	4	8	5	5
G	9	5	7	7	3	4	2	5	6	6	5	5	4	4	5	5	6	5	0
C	7	7	5	5	9	4	8	3	3	9	4	4	7	4	1	10	2	4	7
T	2	4	4	6	7	3	7	6	7	0	3	5	4	7	4	2	5	7	9

1E. 503 gene Effective sequences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	101	259	271	192	236	252	211	252	230	198	220	246	225	243	234	195	247	264	266
G	481	235	197	247	226	205	263	218	194	227	242	210	222	233	207	249	205	209	176
C	181	205	177	253	213	182	229	199	209	233	217	199	210	191	171	216	205	154	204
T	97	161	215	168	185	221	157	191	227	202	181	205	203	193	248	200	203	231	205

2.2 Effectiveness definition for individual guidelines

As G1, G2 and G3 specify preferred/effective and unpreferred/ineffective nucleotides for selecting effective gene-silencing sequences, it might be possible to estimate the effectiveness of the individual guidelines by comparing the occurrence probability of nucleotides specified by the guidelines with the corresponding nucleotide average probability in their sequence populations. That is, their effectiveness can be evaluated by comparing the probability of the nucleotides assigned by individual guidelines with the occurrence probabilities of the corresponding nucleotides averaged over the entire target sequence population.

For this evaluation we used the ratio ***ER*** defined as follows:

$$\boldsymbol{ER} = \frac{\boldsymbol{NR}_p}{\boldsymbol{AN}_i} \tag{1}$$

where $\boldsymbol{NR}_p$ is the nucleotide probability assigned by the guideline at position p (*e.g.*, $\boldsymbol{NR}_3 = 0.28$) and $\boldsymbol{AN}_i$ is the occurrence probability of the corresponding nucleotide i (A, G, C, or T) averaged over the entire target sequence population (*e.g.*, $\boldsymbol{AN}_A = 0.26$).

Suppose, for example, that the nucleotide A at position 3 is specified as the preferred one by some guideline and there are 200 effective siRNA sequences as the entire target sequence population. These sequences therefore have 200 nucleotides at each position from 1 to 19 and have 3800 (=200x19) nucleotides in the population. If the numbers of individual nucleotide occurrences at position 3 are, for example, A=56, C=48, G=62 and T=34 and those of the entire population are A=988, C=912, G=1102 and T=798, $\boldsymbol{NR}_3$, $\boldsymbol{AN}_A$ and ***ER*** are respectively computed as 0.28 (=56/200), 0.26 (=988/3800) and 1.08 (=0.28/0.26). This implies that the nucleotide A at position 3 indicates the average frequency level although it is specified as the preferred nucleotide by the guideline. The ratio ***ER*** therefore indicates the effectiveness of the individual guideline. If the ***ER*** for the preferred/effective nucleotide is markedly larger than 1, the nucleotide specified by that guideline is effective for other genes. And if the ***ER*** for the specified nucleotide is markedly lower than 1, the guideline is ineffective for other genes. A converse relation applies to the ***ER***s for the unpreferred/ineffective nucleotides: the guideline is ineffective for other genes when the ***ER*** > 1 and is effective when the ***ER*** < 1.

Although it is generally difficult to determine the exact ***ER*** needed to distinguish effective guidelines from ineffective guidelines, we regarded ***ER*** > 1.25 to indicate an effective nucleotide and ***ER*** < 0.75 to indicate an ineffective nucleotide. The reason for choosing an ***ER*** at least 25%

greater than 1 as a preferred value and an ***ER*** at least 25% less than 1 as an unpreferred one is that we expected these values to provide 99% statistical significance for the population of more than 300 sequences. Similarly, for the unpreferred/ineffective nucleotide, we regarded ***ER*** < 0.75 as effective and ***ER*** > 1.25 as ineffective.

2.3 Statistical significance analysis

If a gene-silencing phenomenon depends on siRNA sequences, nucleotide occurrence features should be evident when a large number of effective target sequences are analysed statistically. That is, it is possible to test the significance of each nucleotide at individual positions (*i.e.*, 1 to 19) on the basis of the occurrence probabilities of each nucleotide at individual positions and in the entire target sequence population. We used the following significance testing to clarify individual nucleotide positional dependencies:

$$Z = \frac{\left| p_a - p_b \right|}{\sqrt{P(1-P)(1/n_a + 1/n_b)}}, \tag{2}$$

where P_a is the probability of each nucleotide occurring at the individual sequence sites (*i.e.*, positions 1 to 19), P_b is the occurrence probability of each nucleotide averaged over the entire target sequence population, P is P_a and P_b arithmetic means, n_a is the number of nucleotides at individual positions (sites) and n_b is the total number of nucleotides in all positions (sites) [21] [22].

Suppose, for example, that we have 200 effective siRNA sequences. If the occurrence probability of the nucleotide G at position 7 is 0.35 (70/200) and the occurrence probability of G in the entire target sequence population is 0.28 (1064/(200×19)), the 95% significance probability of the nucleotide G at position 7 would be indicated by a *z* value of 2.14.

Statistical significant nucleotide selection

As the two-sided statistical test has two types of significance values, higher (upper) and lower levels of significance, they are expressed as follows:

Higher-significance nucleotide ($\boldsymbol{HN}_p^v$) and
Lower-significance nucleotide ($\boldsymbol{LN}_p^v$),
where *H* denotes higher, *L* denotes lower and *N* is a nucleotide,

v : 95 – significance probability is 95% (level of significance = 0.05),
99 – significance probability is 99% (level of significance = 0.01),
p : nucleotide position (site) (*i.e.*, 1–19).

3. Results and Discussion

3.1 Analysis I: Qualitative analysis of three guidelines

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 trail-and-error method [7], 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 larger the number of candidates becomes, the more difficult it is to get gene-silencing candidates.

Reynolds *et al*. recently analysed 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. The preferred nucleotides (G1) for effective siRNA designs are shown in Figure 1. Ui-tei *et al*. and Amarzguioui & Prydz also reported guidelines and an algorithm for effective siRNA designs based on their literature. The corresponding effective and ineffective nucleotides for siRNA designs are also summarized in Figure 1.

	position	1	3	6	10	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

G1: Reynolds et al. G2: Ui-Tei et al. G3: Amarzguioui et al.

Figure 1. Effective and ineffective nucleotides specified in the individual guidelines.

Position: nucleotide position from 1 to 19 (5' to 3', cDNA form). Preferred: effective, unpreferred: ineffective. G1: Reynolds *et al*.; eight criteria: (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. G2: Ui-tei *et al*.; four rules: (1) A or T effective and G or C ineffective at position 19, (2) G or C effective and A or T ineffective 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. G3: Amarzguioui, M. & Prydz, H; six rules: (1) G or C positive and T negative at position 1, (2) A positive at position 6, (3) T negative at position 10, (4) T positive at position 13, (5) C positive at position 16, (6) A or T positive and G negative at position 19.

What is a good guideline for selecting gene-silencing siRNA? It is one that is effective for selecting siRNA sequences causing gene degradation and that can be applied for many genes. To determine the effectiveness of the recent reported guidelines, we first clarified the relations between them from a qualitative viewpoint and then quantitatively evaluated their effectiveness for other genes.

From Figure 1 we obtained the following consistencies among the guidelines reported by Reynolds *et al*., and Ui-tei *et al*. and Amarzguioui & Prydz [22].

1) Consistency among three guidelines (G1, G2 and G3):
   A or T effective at position 19
2) Consistency between pairs of guidelines (G1 and G3, G2 and G3):
   T effective at position 13 (G1 and G3)
   G or C effective at position 1 (G2 and G3)
   T ineffective at position 1 and G ineffective at position 19 (G2 and G3)

Because we found only these few consistencies, we thought it would be difficult use these guidelines to select effective target sequences. If only these few nucleotide consistencies were used for selecting target sequences, many sequence candidates would be extracted from the target genes and it would be hard to select a few final candidates for synthesizing siRNAs.

3.2 Analysis II: Quantitative individual guideline effectiveness for other genes

We first examined the effectiveness of the individual guidelines (G1, G2 and G3) for the reported genes (see Materials and Methods). If G1, G2 and G3 were effective for selecting effective siRNA sequences for other genes, the preferred and unpreferred nucleotides for effective gene-silencing sequences would show the similar occurrence tendencies with each of the guidelines. This can be determined by analyzing the ***ER*** values obtained when the individual guidelines are applied for other genes.

3.2.1 Guideline 1 (G1) effectiveness for selecting sequences effective for other genes

In G1 there are four preferred nucleotides (I, II, III and IV) (see Materials and Mehtods). If G1 were applicable to designing siRNAs for other genes (MG2 and MG3), the four nucleotides in the effective sequences would occur at the specified sites more often than they would occur elsewhere. That is, if G1 were effective for the other genes, the four nucleotides should be higher than the G1 property in Figure 2(A). The relations between G1 and the reported genes MG2 and MG3 are shown in Figure 2(A), which shows that G1 indicates many reverse tendencies (under the G1 property) for effective sequences in other genes and only nucleotide IV demonstrates G1 effectiveness for MG2 and MG3. In other words, it could be inferred that it is difficult to use G1 for selecting sequences effective in other genes.

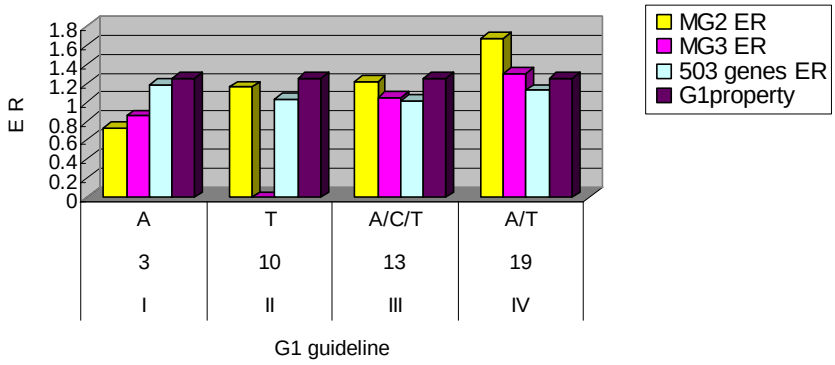
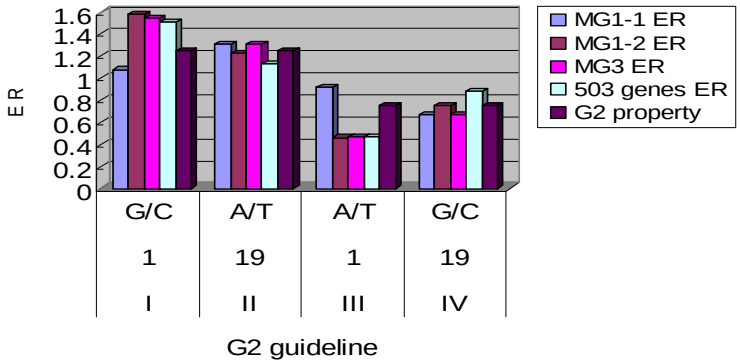


Figure 2. Individual guideline effectiveness for other genes.

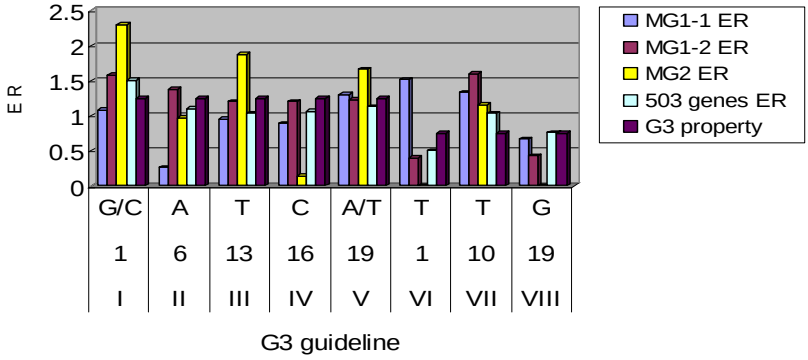
2(A) Effectiveness of Guideline 1 (G1) for selecting sequences effective for other genes.

I (A at position 3), II (T at position 10), III (A, C or T at position 13) and IV (A or T at position 19). ***ER***: Equation (1). G1 property: ***ER***=1.25 (see Materials and Methods) for I, II, III and IV.

a) In the case of effective sequences for MG2, the nucleotide IV (***ER***=1.67) indicates a clear G1 property but I (***ER***=0.73) shows a reverse tendency, whereas the nucleotide III (***ER***=1.22) shows nearly the same G1 property. b) For MG3, only the nucleotide IV (***ER***=1.31) indicates the G1 property tendency. In contrast the nucleotides I (***ER***=0.86) and II (***ER***=0) show reverse tendencies. Especially, II indicates a distinct difference from G1 property. c) For 503 cDNAs, the nucleotide I (***ER***=1.19) indicates closely some G1 property but II (***ER***=1.04), III (***ER***=1.02) and IV (***ER***=1.13) show no G1 property (all ERs are average levels).



2(B) Effectiveness of Guideline 2 (G2) for selecting sequences effective for other genes.
I (G or C at position 1), II (A or T at position 19), III (A or T at position 1) and IV (G or C at position 19). G2 property: ***ER***=1.25 for I and II and ***ER***=0.75 (see Materials and Methods) for III and IV.
a) In the case of effective sequences for MG1-1, the nucleotides II (***ER***=1.31) and IV (***ER***=0.67) indicate G2 property tendencies. b) For MG1-2, the nucleotides I (***ER***=1.58), II (***ER***=1.23), III (***ER***=0.45) and IV (***ER***=0.75) indicate G2 property tendencies. c) For MG3, the nucleotides I (***ER***=1.55), II (***ER***=1.31), III (***ER***=0.47) and IV (***ER***=0.67) also indicate G2 property tendencies. d) For 503 cDNAs, the nucleotides I (***ER***=1.51) and III (***ER***=0.47) indicate a clear G2 property, whereas II (***ER***=1.13) and IV (***ER***=0.88) show no G2 property.



2(C) Effectiveness of Guideline 3 (G3) for selecting sequences effective for other genes.
I (G or C at position 1), II (A at position 6), III (T at position 13), VI (C at position 16), V (A or T at position 19), VI (T at position 1), VII (T at position 10), VIII (G at position 19). G3 property: ***ER***=1.25 for from I to V and ***ER***=0.75 for VI, VII and VIII.
a) In the case of effective sequences for MG1-1, the nucleotides V (***ER***=1.31) and VIII (***ER***=0.67) indicate G3 property tendencies. In contrast II (***ER***=0.26) and VI (***ER***=1.52) show distinct reverse tendencies and IV (***ER***=0.89) and VII (***ER***=1.33) also show reverse tendencies. b) For MG1-2, all nucleotides except VII (***ER***=1.6) indicate a G3 property and only VII shows a distinct reverse tendency. c) For MG2, the nucleotides I (***ER***=2.3), III (***ER***=1.88), V (***ER***=1.67), VI (***ER***=0) and VIII (***ER***=0) indicate a strong G3 property. In contrast, the nucleotides IV (***ER***=0.14) and VII (***ER***=1.16) indicate a distinct reverse tendency. d) For 503 cDNAs, the nucleotides I (***ER***=1.51) and VI (***ER***=0.5) indicate a clear G3 property and VIII (***ER***=0.76) shows nearly the same G3 property, whereas II (***ER***=1.1), III (***ER***=1.06), IV (***ER***=1.08), V (***ER***=1.08) and VII (***ER***=0.98) demonstrate no G3 property (average levels).

3.2.2 Guideline 2 (G2) effectiveness for selecting sequences effective for other genes

In G2 there are two preferred nucleotides (I and II) and two unpreferred nucleotides (III and IV) (see Materials and Methods). If G2 were effective for other genes, the same occurrence tendencies would be expected there. That is, the nucleotides I and II should be higher than the G2 property, the nucleotides III and IV should be lower than the G2 property in Figure 2(B). The

relations between G2 and the reported genes MG1-1, MG1-2 and MG3 are shown in Figure 2(B). Overall, it could be inferred that G2 has similar tendencies for the other gene effective sequences except MG1-1. However, as both the nucleotides I and III are located at position 1 and both the nucleotides II and IV are located at position 19, there are many possible candidate sequences between positions 2 and 18. So even though candidate sequences satisfy G2, there might be many alternative sequences. This means that the selection of effective sequences becomes a time-consuming task.

3.2.3 Guideline 3 (G3) effectiveness for selecting sequences effective for other genes

In G3 there are five preferred nucleotides (I, II, III, IV and V) and three unpreferred nucleotides (VI, VII and VIII) (see Materials and Methods). If G3 were effective for other genes, the eight nucleotides would show the same occurrence tendencies there. That is, the nucleotides I to V should be higher than the G3 property and the nucleotides VI to VIII should be lower than the G3 property in Figure 2(C). The relations between G3 and MG1-1, MG1-2 and MG2 are shown in Figure 2(C). Overall, G3 has similar tendencies for the other genes except MG1-1.

We then analysed the individual guideline effectiveness for 860 effective siRNA sequences from 503 different mammalian cDNAs in the literature.

3.2.4 Individual guideline (G1, G2 and G3) effectiveness for selected effective siRNA sequences in 503 genes

ER*s* indicating the degree of G1 effectiveness for 860 sequences from 503 cDNAs are also shown in Figure 2(A). As the range of ***ER*** for 860 sequences is from 1.02 to 1.19, there is no nucleotide for which ***ER*** > 1.25. The results thus show that G1 provides no effective guidance for selecting gene-silencing sequences for many mammalian genes.

ER*s* indicating G2 and G3 effectiveness for 860 sequences are respectively shown in Figures 2(B) and 2(C). G2 shows that the nucleotides I and III are effective indications for 860 sequences. As both I and III are located at position 1, however, provides effective guidance only at position 1. G3 also indicates the same tendency. That is, the nucleotides I (***ER***=1.46) and VI (***ER***=0.5) provide effective guidance only at position 1. As G1 is not effective and G2 and G3 are effective only at position 1 for 860 sequences from 503 cDNAs, these three guidelines may not be effective for many mammalian genes.

3.3 Individual guideline efficacies

For clarity we use positive (+) and negative (-) indications of positional features for the individual guidelines G1, G2 and G3 shown in Figure 3. Although G1 has two nucleotides showing positive indications (*i.e.*, III for MG2 and IV for MG2 and MG3), there are two nucleotides showing negative indications (*i.e.*, II for MG3 and I for MG2 and MG3). In addition, there is no G1 tendency for 503 genes. This indicates that G1 is partially useful for specific genes. On the other hand, as G2 has four nucleotides showing positive tendencies (*i.e.*, the nucleotides I, II, III and IV for MG1-2 and MG3, and II and IV for MG1-1, and I and III for 503 genes), it indicates effective tendencies for MG1-2 and MG3. As the effective nucleotides I and III for 503 genes are located at position 1, however, there is no effective nucleotide indicated for any of the other positions. This might imply that there are many candidates satisfying the nucleotide condition specified at position 1. In contrast, as G3 has seven nucleotides showing positive tendencies (*i.e.*, the nucleotides V and VIII for MG1-1, I to VI and VIII for MG1-2, I, III, V, VI and VIII for MG2, I, II, VI and VIII for 503 genes), it provides effective indications for MG1-2 and MG2 but ineffective indications for

MG1-1. However, although G3 has four effective nucleotides (4/8) for 503 genes, only two (I and VI) satisfies with G3 property and their positions are located at 1. So it might be possible to select many candidates satisfying these conditions. It may therefore be hard to determine whether they are effective gene-silencing candidates.

3(A)

	I	II	III	IV
MG2	-		+	+ +
MG3	-	- -		+
503 genes				

3(B)

	I	II	III	IV
MG1-1		+		+
MG1-2	+ +	+	+ +	+
MG3	+ +	+	+ +	+
503 genes	+ +		+ +	

3(C)

	I	II	III	IV	V	VI	VII	VIII
MG1-1		- -		-	+	- -	-	+
MG1-2	+ +	+	+	+	+	+ +	- -	+ +
MG2	+ +		+ +	- -	+ +	+ +	-	+ +
503 genes	+ +	+				+ +		+

Figure 3. Individual guideline effectiveness clarification.

ER for individual preferred nucleotides is classified into the following categories: $1.45 \leq ER$: strong tendency "+ +"; $1.2 \leq ER < 1.45$:same tendency "+"; $0.9 \leq ER < 1.2$: no tendency; $0.65 \leq ER < 0.9$: reverse tendency "-"; $ER < 0.65$:distinct reverse tendency "- -".***ER*** for individual unpreferred nucleotides is classified into the following categories: $ER \leq 0.55$: strong tendency "+ +"; $0.55 < ER < 0.8$:same tendency "+"; $0.8 \leq ER < 1.0$:no tendency; $1.0 \leq ER < 1.45$:reverse tendency "-"; $1.45 \leq ER$: distinct reverse tendency "- -".

3(A) Guideline G1 effectiveness. 3(B) Guideline G2 effectiveness. 3(C) Guideline G3 effectiveness.

3.4 siRNA sequence selection problems using the previous guidelines

Our qualitative and quantitative analyses revealed that G1, G2 and G3 are not always effective selection rules for other mammalian genes. In other words, though some rules from the guidelines are suitable for getting effective sequences for specific genes, they might sometimes be unsuitable for selecting sequences for other genes. Since the individual guidelines G1, G2 and G3 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 G1, G2 and G3. As a result, many candidate sequences might be selected. Experimentally evaluating whether the selected sequences provide effective gene degradation, however, is a costly and time-consuming task.

3.5 The proposed measure based on the positional features of significant nucleotides

Since it could be inferred that the previous guideline problems were based on specific gene analyses, we examined whether there are nucleotide occurrence specificities for many effective siRNA sequences reported in the literature (see Materials and Methods).

We analysed 860 sequences listed in Table 1E by using Equation (2) and obtained many higher-significance and lower-significance nucleotides shown in Table 2. They are mostly different from the previous reported nucleotides obtained by using G1, G2 and G3.

Table 2. Higher-significance and lower-significance nucleotides and their static values.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Sig. (Z)	**G** (18.1)	**A** (2.3)	**A** (3.2)	**C** (3.9)		**T** (2.1)	**G,C** (2.2,2)		**T** (2.6)	**C** (2.3)					**T** (4.2)			**A,T** (2.7, 2.9)	**A** (3)
Sig. (Z)	**A,T** (9.7, 7.8)	**T** (2.7)	**G,C** (2.8,2)	**A,T** (2.8, 2.1)		**G** (2.2)	**T** (2.9)		**G** (3)	**A** (2.3)					**G,C** (2,2.5)	**A** (2.5)	**G** (2.2)	**C** (3.9)	**G** (4.2)

Gene silencing measure (score) definition

A target siRNA sequence including many higher-significance nucleotides and a few lower-significance nucleotides could be inferred to become a highly effective gene-silencer. A measure (priority score ***S***) based on this idea is therefore defined in the following way.

$$S = a\left(\sum HN_p^v\right) - b\left(\sum LN_p^v\right) \tag{3}$$

where ***a*** and ***b*** are weighting factors for higher and lower scores, normally ***a***=***b***=1.

HN_p^v and LN_p^v are respectively higher-significance and lower-significance nucleotides (see Materials and Methods).

Equation (3) shows that the larger ***S*** becomes, the greater the likelihood that the sequence is effective for gene-silencing.

3.6 Effectiveness of the proposed score method

The purpose of the score setting is to indicate the sequence priority for selecting new siRNA candidates. This is because it is necessary to select several of the highest-ranked sequences as target sequences for dsRNA syntheses. From this point of view, we evaluated the effectiveness of the proposed score for effective and ineffective siRNA sequences of MG1-1, MG1-2, MG2 and MG3 by using equation (3) and Table 2. Since there were ups and downs in the computed scores of the individual sequence classes, we calculated the averages for them. The average scores of the effective sequences for MG1-1, MG1-2, MG2 and MG3 were respectively 2.2, 8.8, 13.2 and 5.9, whereas those of the ineffective sequences were -1.5, -9.5, -14.3 and -3.8. These scores therefore reveal that the proposed method might be useful for selections of effective siRNA candidates. We also examined the average scores obtained by the previously reported score methods of Reynolds et al., Ui-Tei et al., Amarzguioui and Prydz, and Hsieh et al. [17]. The relative relations between scores of the previous methods and those of the proposed method are shown in Figure 4. The results indicate that the previous methods are not always clear correspondences between the scores and the effective and ineffective siRNA sequences. The methods of Reynolds et al. and Hsieh et al., for

example, show positive values for both effective and ineffective siRNAs of MG1-1, MG1-2 and MG3, and don't indicate distinct score differences between the effective and ineffective siRNAs. In addition, although the methods of Ui-Tei et al. and Amarzguioui and Prydz indicate the correspondences between the individual average scores and the effective and ineffective siRNAs for MG1-2, MG2 and MG3, the relative score differences between the effective and ineffective siRNAs are not so big as shown in Figure 4. In addition, the maximum and minimum score ranges of their methods are respectively from 2 to -2 and from 8 to -2. [17]. These score ranges imply that the restricted discrete scores might be assigned to the candidate sequences. For example, the method of Ui-Tei et al. may assign five kinds of scores, i.e., (2, 1, 0, -1, -2) to the candidate sequences. This indicates that there might be many same score sequences and the difficulty of selecting several candidates. Therefore, these previous method results imply that it is difficult to assign the priority of new siRNA candidates according to the obtained scores. On the other hand, as the range of the proposed score is 46.8 to -43.7 from Table 2, it is easy to distinguish the priority for the candidate sequences. The proposed scores, for example, indicate clear correspondences for the effective and ineffective siRNAs of MG1-2, MG2 and MG3. This therefore implies that the proposed score can easily be used for selecting high-potential siRNA candidates.

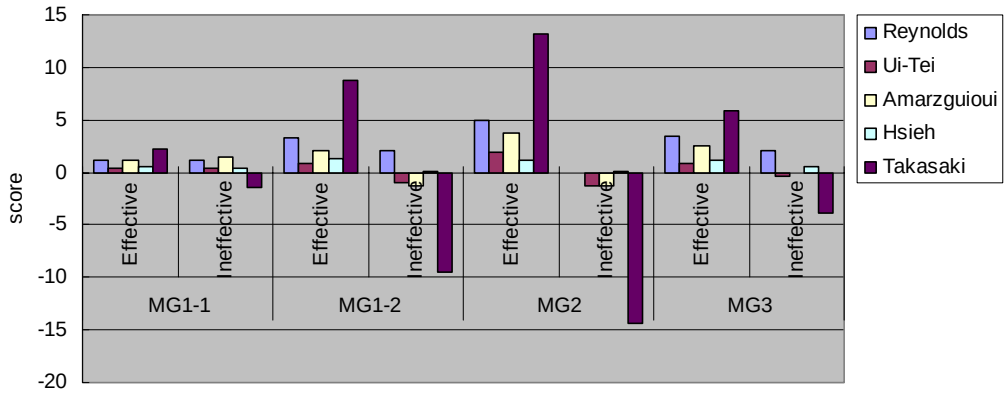


Figure 4. Score comparisons between the proposed method and other methods.
Scores of the effective and ineffective siRNA sequences were computed on the basis of positional scores of the individual reported guidelines shown in Saetrom et al. 2004[17].

3.7 Individual guideline effectiveness for ineffective siRNA sequences

If G1, G2 and G3 were effective for eliminating ineffective siRNA sequences for other genes, with each of the guidelines the individual preferred and unpreferred nucleotides for ineffective gene-silencing sequences would show the reverse occurrence tendencies. That is, the preferred nucleotide ***ERs*** would be lower than 0.75 for the ineffective sequences, whereas the unpreferred nucleotide ***ERs*** would be higher than 1.25 for them.

3.7.1 G1 effectiveness for eliminating ineffective sequences for other genes

In ineffective sequences for other genes, the nucleotides I, II, III and IV should be have a distinctly lower probability of occurring at the specified sites than other nucleotides have. If G1 indicates the same tendency for other genes, the same occurrence phenomena are expected (*i.e.*, ***ERs*** < 0.75). That is, the nucleotides I to IV should be lower than the G1 property in Figure 5(A). The relations between G1 and MG2 and MG3 are shown in Figure 5(A). The results show that the

ratio of nucleotide ***ER***<0.75 for MG2 is 2/4 (I and IV), whereas that for MG3 is zero. It could therefore be inferred that G1 is not always useful for eliminating ineffective siRNA sequences.

3.7.2 Effectiveness of G2 and G3 for eliminating ineffective sequences for other genes

If G2 indicates the same tendency for other genes, ***ER****s* for the preferred nucleotides I and II are expected to be lower than 0.75 and ***ER****s* for the unpreferred nucleotides III and IV are expected to be higher than 1.25. That is, the nucleotides I and II should be lower than the G2 property and the nucleotides III and IV should be higher than the G2 property in Figure 5(B). The relations between G2 and MG1-1, MG1-2 and MG3 are shown in Figure 5(B). The results show that G2 provides a good guideline for MG1-2 — that is, the ***ER****s* for I and II are less than 0.75 and the ***ER****s* for III and IV are higher than 1.25 — whereas it show distinct reverse tendencies for MG1-1 and slightly reverse tendencies for MG3. This means that G2 is not always useful for eliminating ineffective sequences.

G3 has five preferred nucleotides (I to V) and three unpreferred nucleotides (VI, VII and VIII), and the ***ER****s* for the preferred nucleotides are expected to be lower than 0.75 and the ***ER****s* for the unpreferred nucleotides should be higher than 1.25. That is, the nucleotides I to V should be lower than the G3 property and the nucleotides VI to VIII should be higher than the G3 property in Figure 5(C). The relations between G3 and MG1-1, MG1-2 and MG2 are shown in Figure 5(C). The results show that G3 provides a relatively good guideline for MG1-2 and MG2 (I, II, V, VI and VIII for MG1-2 and I, III, V, VII and VIII for MG2), whereas it shows a negative guideline for MG1-1 (only VII is higher than 1.25).

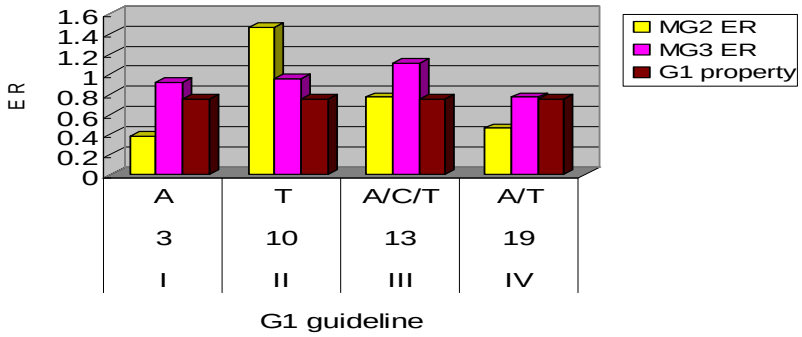
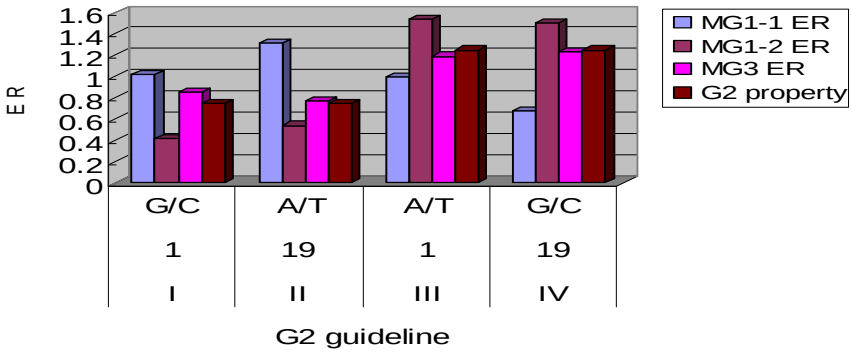


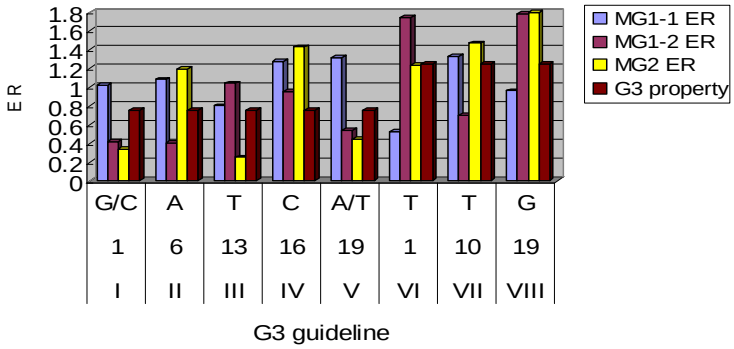
Figure 5. Relative effectiveness individual guideline for eliminating ineffective siRNA sequences.

5(A) Effectiveness of Guideline 1 (G1) for eliminating ineffective sequences for other genes.
In this case, G1 property (***ER***) is set to 0.75 (see Materials and Methods) for I, II, III and IV. a) In the case of ineffective sequences for MG2, the nucleotides I (***ER***=0.38) and IV (***ER***=0.46) indicate a clear G1 property and III (***ER***=0.77) shows nearly the same G1 property, whereas II (***ER***=1.47) demonstrates a distinct reverse tendency. b) For MG3, the nucleotide IV (***ER***=0.77) indicates nearly the same G1 property, whereas I (***ER***=0.92), II (***ER***=0.95) and III (***ER***=1.11) show reverse tendencies.



5(B) Effectiveness of Guideline 2 (G2) for eliminating ineffective sequences for other genes.

In this case, G2 property (***ER***) is set to 0.75 for the nucleotides I and II and to 1.25 (see Materials and Methods) for III and IV. a) In the case of ineffective sequences for MG1-1, as the nucleotides I (***ER***=1.02), II (***ER***=1.31), III (***ER***=1) and IV (***ER***=0.68) are different from the G2 property and indicate reverse tendencies. b) For MG1-2, the nucleotides I (***ER***=0.42), II (***ER***=0.54), III (***ER***=1.54) and IV (***ER***=1.5) indicate a strong G2 property. c) For MG3, only the nucleotide II (***ER***=0.77) indicates what is nearly the G2 property.



5(C) Effectiveness of Guideline 3 (G3) for eliminating ineffective sequences for other genes.

In this case, G3 property (***ER***) is set to 0.75 for nucleotides I through V and to 1.25 for nucleotides VI, VII and VIII. a) In the case of ineffective sequences for MG1-1, the nucleotide VII (***ER***=1.33) indicates a G3 property tendency and III (***ER***=0.8) shows nearly G3 property. In contrast, the nucleotides I (***ER***=1.02), II (***ER***=1.08), IV (***ER***=1.27), V (***ER***=1.31), VI (***ER***=0.53) and VIII (***ER***=0.96) show reverse tendencies. b) For MG1-2, the nucleotides I (***ER***=0.42), II (***ER***=0.41), V (***ER***=0.54), VI (***ER***=1.74) and VIII (***ER***=1.78) indicate a distinct G3 property but III (***ER***=1.04), IV (0.95) and VII (***ER***=0.7) show reverse tendencies. c) For MG2, the nucleotides I (***ER***=0.34), III (***ER***=0.25), V (***ER***=0.45), VII (***ER***=1.47) and VIII (***ER***=1.8) indicate a strong G3 property and VI (***ER***=1.23) tends to show a G3 property. In contract, the nucleotides II (***ER***=1.19) and IV (***ER***=1.43) show distinct reverse tendencies.

3.8 Optimal GC content

We examined the GC content of effective and ineffective siRNA sequences for reported genes. The distribution of GC content for the sets of reported genes is listed in Table 3. The results indicate that there is no big difference between effective and ineffective sequences. It could therefore be inferred that gene-silencing effectiveness does not depend on GC content. As the average GC content for 860 effective sequences is 50.8% with a standard deviation of 8.9%, this value could be used as a guideline.

Table 3. GC content distribution.
Effect seqs: Effective sequences (%), ( ): standard deviation, Ineffective seqs: Ineffective sequences

	MG1-1	MG1-2	MG2	MG3	860 seqs
Effect seqs	48.4(7)	48.2(11.8)	42.1(8.5)	49.1(13.3)	50.8(8.9)
Ineffective seqs	47.6(8.2)	46.1(10)	62.2(9.6)	52.4(13.3)	

4. Conclusions

Analyzing the reported siRNA design guidelines from qualitative and quantitative points of views, we found that they were not always effective selection rules for many other mammalian genes. Though some rules from the guidelines are suitable for extracting effective sequences for specific genes, they might sometimes be unsuitable for selecting sequences for other genes. Since the gene-silencing efficacy depends very much on the target sequence positions selected from the target gene, we examined 860 effective siRNA sequences from 503 different mammalian cDNAs in the literature. As a result, we got many preferred and unpreferred nucleotides different from the ones used in the previous guidelines. We proposed the gene silencing measure based on the positional features of significant nucleotides and demonstrated the effectiveness of the proposed measure compared with the recently reported other scoring methods. In this paper we also discussed the elimination of ineffective siRNA sequences from target candidates and optimal GC content in siRNA sequences.

References

[1] M. Amarzguioui and H. Prydz, An algorithm for selection of functional siRNA sequences, *Biochem. Biophys. Res. Commun.*, **316**, 1050-1058 (2004).

[2] A.M. Chalk, C. Wahlestedt and E.L.L. Sonnhammer, Improved and automated prediction of effective siRNA, *Biochem. Biophys. Res. Commun.*, **319**, 264-274 (2004).

[3] D.M. Dykxhoorn, C.D. Navia, and P.A. Sharp, Killing the messenger: Short RNAs that silence gene expression, *Nature Review*, **4**, 457-467 (2003).

[4] S.M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber *et al.*, Duplexes of 21-nucleotide RNAs mediate RNA interference in mammalian cell culture, *Nature*, **411**, 494-498 (2001).

[5] S.M. Elbashir, W. Lendeckel, and T. Tuschl, RNA interference is mediated by 21- and 22-nucleotide RNAs, *Genes Dev.*, **15**, 188-200 (2001).

[6] S.M. Elbashir, J. Martinez, A. Patkaniowska, W. Lendeckel, and T. Tuschl, Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysates, *EMBO J.*, **20**, 6877-6888 (2001).

[7] S.M. Elbashir, J. Harborth, K. Weber, and T. Tuschl, Analysis of gene function in somatic mammalian cells using small interfering RNAs, *Methods*, **26**, 199-213 (2002).

[8] A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver *et al.*, Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*, *Nature*, **391**, 806-811 (1998).

[9] G.J.Hannon, RNA interference, *Nature*, **418**, 244-251 (2002).
[10] T. Holen, M. Amarzguioui, M.T. Wiiger, E. Babaie and H. Prydz, Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor, *Nucleic Acids Res*., **30**, 1757-1766 (2002).
[11] A.C. Hsieh, R. Bo, J. Monola, F. Vazquez, O. Bare et al., 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 (2004).
[12] A. Khvorova, A. Reynolds, S.D. Jayasena, Functional siRNAs and miRNAs exhibit strand bias, *Cell*, **115**, 209-216 (2003).
[13] V. Mittal, Improving the efficiency of RNA interference in mammals, *Nature Rev. Genetics*, **5**, 355-365 (2004).
[14] Y. Naito, T. Yamada, K. Ui-Tei, S. Morishita and K. Saigo, siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference, *Nucleic Acids Res*., **32**, W124-W129 (2004).
[15] A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W.S. Marshall et al., Rational siRNA design for RNA interference, *Nat. Biotech*., **22**, 326-330 (2004).
[16] J. Santoyo, J.M. Vaguerizas and J. Dapozo, Highly specific and accurate selection of siRNAs for high-throughput functional assays, *Bioinfomatics*, **21**, 1376-1382 (2005).
[17] P. Saetrom and O.Jr. Snove, A comparison of siRNA efficacy predictors, *Biochem. Biophy. Res. Commun.*, **321**, 247-253 (2004).
[18] D.S. Schwarz, G. Hutvagner, T. Du, Z. Xu, N. Aronin et al., Asymmetry in the assembly of the RNAi enzyme complex, *Cell*, **115**, 199-208 (2003).
[19] P.A. Sharp, RNA interference-2001, *Genes Dev.*, **15**, 485-490. (2001).
[20] O.Jr. Snove, M. Nedland, S.H. Fjeldstad, H. Humberset, O.R. Birkeland et al., Designing effective siRNAs with off-target control, *Biochem. Biophy. Res. Commun.*, **325**, 769-773 (2004).
[21] S. Takasaki, S. Kotani and A. Konagaya, An effective method for selecting siRNA target sequences in mammalian cells, *Cell Cycle*, **3**, 790-795 (2004).
[22] S. Takasaki, S. Kotani and A. Konagaya, Selecting Effective siRNA Target Sequences for Mammalian Genes, *RNA Biology*, **2**, 21-27 (2005).
[23] M. Truss, M. Swat, S.M. Kielbasa, R. Schafer, H. Herzed et al., 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 (2005).
[24] K. Ui-Tei, Y. Naito, F. Takahashi, T. Haraguchi, H. Ohki-Hamazaki et al., Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference, *Nucleic Acids Res*., **32**, 936-948 (2004).