

# A Three-Phase Algorithm for Computer Aided siRNA Design

Hong Zhou  
Saint Joseph College, West Hartford, CT 06117, USA  
hzhou@sjc.edu

Xiao Zeng  
Superarray Bioscience Corporation, 7320 Executive Way, Frederick, MD 21704, USA  
xzeng@superarray.net

Yufang Wang and Benjamin Ray Seyfarth  
University of Southern Mississippi, Hattiesburg, MS 39406, USA

**Keywords:** siRNA, RNA interference, three-phase, Smith-Waterman, BLAST

**Received:** July 10, 2005

*As our knowledge of RNA interference accumulates, it is desirable to incorporate as many selection rules as possible into a computer-aided siRNA-designing tool. This paper presents an algorithm for siRNA selection in which nearly all published siRNA-designing rules are categorized into three groups and applied in three phases according to their identified impact on siRNA function. This tool provides users with the maximum flexibility to adjust each rule and reorganize them in the three phases based on users' own preferences and/or empirical data. When the generally accepted stringency was set to select siRNA for 23,484 human genes represented in the RefSeq Database (NCBI, human genome build 35.1), we found 1,915 protein-coding genes (8.2%) for which none suitable siRNA sequences can be found. Curiously, among these 1,915 genes, two had validated siRNA sequences published. After close examination of another 105 published human siRNA sequences, we conclude that (A) many of the published siRNA sequences may not be the best for their target genes; (B) some of the published siRNA may risk off-target silencing; and (C) some published rules have to be compromised in order to select a testable siRNA sequence for the hard-to-design genes.*

*Povzetek: Predstavljen je algoritem za obdelovanje genoma.*

## 1 Introduction

Since the seminal paper published by Craig C. Mello's group in 1998 [1], RNA interference (RNAi) has emerged as a powerful technique to knock out/down the expression of target genes for gene function studies in various organisms [2,3,4]. What is truly remarkable about the RNAi effect is that it is sequence-specific. This means that as long as we know the sequence of the transcript to be targeted, we can design a short double-stranded RNA (small interfering RNA or siRNA) to knock down, if not eliminate the expression of the target gene without changing the genetic make-up of the cells. Compared to the anti-sense oligonucleotide technology developed earlier [5,6], RNAi is much more effective because RNAi is achieved by catalytic components within the cell [1,7,8,9].

Understandably, how to design the best siRNA has become an intense competition between academic research groups as well as commercial providers of siRNA. The following is a summary of some major designing rules published.

- **The length of functional siRNAs:** The length of siRNA ranges from 19 to 30 base pairs (bps) [2,10,11]. Double stranded RNA longer than 30 bps

is likely to invoke an antiviral interferon response, a general shut-down of the cellular translation instead of gene-specific RNAi [12,13,14].

- **The GC content of functional siRNA:** The optimal GC content of siRNA should be between 30% and 55% [10,14,15]. GC-rich sequences, in general, have the tendency to form quadruplex or hairpin structures [16]. Sequences with GC stretches over 7 in a row may form duplexes too stable to be unwound [16,17,18,19]. On the other hand, sequences with extremely low GC content cannot form stable siRNA duplexes.
- **The thermo-stability bias at the 5' end of the antisense strand:** Since it is desirable to have only the antisense strand incorporated into the RISC complex, lowering the thermo-stability at the 5' end of the antisense strand can promote helicase unwind siRNA duplexes from this end [17,20,21].
- **Concerning tandem repeats and palindromes:** Since sequences containing tandem repeats or palindromes may form internal fold-back structures, it is best to avoid any internal repeats or palindromes in the designed siRNA sequence [10]. For the same reason and other concerns [22] [23], long single nucleotide repeats (such as AAAA, UUUU, CCCC or GGGG) should also be avoided [19,24].

Regarding the specific nucleotide positions in siRNA, it has been proposed that base U at position 10, base A at position three, and a base other than G at position thirteen were preferred [10]. However, those experiments were conducted with siRNAs 19 bps in length, it is unknown if the same rules apply to longer siRNAs. While some siRNA design algorithms prefer having the siRNA sequence start with AA [14,24,25], others have pointed out that this rule may result in frequent misses of effective siRNA sequences [17]. Besides, starting with AA may sometimes conflict with the notion that 5' antisense end should be thermodynamically less stable than the 5'-sense end [17,20,21]. It is not clear whether siRNA should be picked within the coding region (CDS) only, though it has been suggested that 5' and 3' untranslated region (UTR) should be avoided [24,25]. However, a recent report showed that targeting 3'-UTR was as efficient as targeting the CDS [26]. If the siRNA (or shRNA, small hairpin RNA) is generated via T7 RNA polymerase, additional rules may apply [27].

While it is desirable to incorporate all of the selection rules into a computer aided siRNA design tool, the complication at the moment is how to rank those published rules, especially when some of the rules are contradictive. Currently, quite a few computer aided siRNA design tools have been published [17,18,19,24,25,27,28,29] and some of those have been made accessible through websites. However, none of those tools has successfully incorporated all the rules above, and most of them treat their employed rules without much differentiation. In general, the existing tools adopt a set of rules and assign each rule an equal or different score, and each siRNA sequence is scored against every rule and only those sequences scoring above a predefined point are selected as valid siRNA sequences. Such a simple selection procedure does not accommodate the possibility that some rules are critical for the validity of a siRNA sequence (must be met), while some rules can only affect the efficiency of the siRNA sequence. Meanwhile, those web-based tools only provide users very limited flexibility, and users cannot reorganize the selection rules based on their own preferences or recent research data.

Although the actual mechanism of which is still unclear, the off-target effect [30] of siRNA is largely attributed to partial sequence homology between siRNA and its unintended targets [31,32]. Most available siRNA design tools use BLAST [33] to filter out siRNA candidates that may cause off-target effect. However, BLAST may overlook significant sequence homologies [17,34]. As an alternative, the Smith-Waterman search algorithm [35] has been proposed to identify all possible off-target sequences [17]. Unfortunately, Smith-Waterman search against the whole-transcriptome is very time-consuming.

This paper presents a three-phase siRNA selection algorithm that can successfully incorporate all the major rules mentioned above effectively in a way that allows the user to optimize the selection process based on their experimental data. The incorporation of the validated rules ensures the effectiveness and specificity of the

selected siRNA sequences. Meanwhile, knowing that some of the rules may not be compatible under certain conditions, this software package has also incorporated maximum flexibility for the users to adjust the selection process based on their own experiment results or their own preferences.

## 2 Materials and methods

### 2.1 Sequence Data

Complete collection of human mRNAs in the NCBI RefSeq database (human genome build 35.1) was used as the experiment dataset. In addition, 107 published siRNA sequences that targeted human genes were collected from prestigious publications.

### 2.2 The Three-Phase Algorithm

The key concept of the three-phase algorithm is to arrange all the necessary siRNA selection rules in three groups of filters according to their impacts on the siRNA efficacy and apply them to the design process in three steps. Each filter represents a specific design rule. Based on the expediency of each rule, the corresponding filter may be assigned the following properties:

- Enabled. If a filter is enabled, it is applied in the selection process; otherwise it is not used at all.
- Mandatory. If a filter is enabled and designated as mandatory, failure to satisfy the rule results in the elimination of the tested siRNA sequence.
- Selective. If a filter is enabled but not designated as mandatory, it is a selective filter by default. siRNA sequences will proceed to the next filter even though they fail to satisfy a “selective” filter.
- Optional. If the validity of a selective filter is yet to be demonstrated, it will be designated as optional.
- Gain. Positive point(s) assigned when a selective/optional filter is satisfied.
- Penalty. Negative point(s) assessed if a selective/optional filter is not met.

As expected, all Phase I filters are mandatory if enabled, eliminating all the sequences containing the most damaging elements for a functional siRNA. All Phase II filters are selective, and will rank eligible siRNA sequences by a final score with the sum of gain and penalty points. Phase III filters represent those rules whose impact on the siRNA functionality has yet to be elucidated and therefore considered optional. The final scores of optional filters will be recorded separately and will not be used to rank the siRNA sequences as with the Phase II filters. Based on the known selection rules, here are 15 filters tested in this work:

Phase I Filters (*by default enabled and mandatory*):

1. The filter for siRNA length (f-len). It requires that the length of the siRNA sequences be between 19 bps to 30 bps, inclusive (not counting the 3' two-nucleotides overhangs).

2. The filter for coding region only (f-coding). It requires that the siRNA sequences be selected only inside the coding sequence.
3. The filter for GC content (f-gc). It requires that the GC content of a siRNA sequence lie between 32 – 55 % inclusive.
4. The filter for repeated sequences (f-repeat). It requires that a siRNA sequence have no internal repeated sequence of length  $\geq 4$ .
5. The filter for internal palindrome (f-palindrome). It requires that a siRNA sequence have no internal palindrome sequence of length  $\geq 5$ .
6. The filter for internal GC stretch (f-stretch). It requires that a siRNA sequence have no GC stretch of length  $> 8$ .
7. The filter for untranslated region (UTR) on mRNA (f-UTR). It requires that a siRNA sequence be 100 nucleotides away from the translation start and stop codons.
8. The filter for the polyA, polyU, polyG and polyC (f-poly). It requires that a siRNA sequence have no AAA, UUU, GGG or CCC.
14. The filter for starting with AA (f-aa). This filter is enabled as optional by default. If the 5' end of sense strand of a siRNA sequence starts with AA, add one point as gain. No penalty is assessed otherwise.
15. The filter for specific nucleotide positions (f-pos). This filter is enabled as optional by default. One point is gained if position three (from 5'-end) of the sense strand is A, another one point is gained if position ten is U, but minus one point is assessed as penalty if position thirteen is G.
16. The filter for the melting temperature (Tm) of the siRNA sequence (f-Tm). For this study, this filter is not enabled. This could measure the Tm value of a siRNA sequence, and set an acceptable range for functional siRNAs [10].

Phase II Filters (*by default enabled and selective*):

9. The filter for the  $\Delta G$  (free energy) at the 5'-end of the antisense strand (f-dga). It requires that the  $\Delta G$  at the 5'-end of antisense should be between -3.6 and -7.2. The gain or penalty of this filter is 1 or 0 respectively.
10. The filter for the  $\Delta G$  (free energy) difference between the 5'-end of the sense strand and the 5'-end of the antisense strand (f-dgd). It requires that the  $\Delta G$  difference ( $\Delta G_{diff} = \Delta G_{5-sense} - \Delta G_{5-antisense}$ ) of a siRNA sequence be less than minus one (-1.0). The gain or penalty of this filter is 1 or -1 respectively.
11. The filter for the number of A/U in the 5'-end pentamer of the antisense strand (f-AU). Among the first five nucleotides at the 5' antisense strand, the gain matches the number of A/U nucleotides present, i.e. if there is one A/U nucleotide the gain would be one point, two A/Us will make two points gain, and so on so forth. No penalty is assessed for zero A/U nucleotide present.
12. The filter for the nucleotide composition at the 5'-end of the sense strand (f-ssnt). If the sense strand of a siRNA sequence starts with a G/C, assess one point gain; otherwise assess minus one point penalty. If there are either one or two A/U present between the second and the fifth nucleotide (inclusive), assess one point as gain; otherwise assess minus one point as penalty.
13. The filter for A/U ending (f-endAU). Two points are gained if the 5'-end antisense strand of a siRNA sequence starts with U. One point is gained if the 5'-end antisense strand of a siRNA sequence starts with A. No penalty is assessed if 5'-end antisense strand of a siRNA sequence starts with G or C.

Phase III Filters:

As stated above, Phase I filters are used to eliminate all sequences that bear at least one unwanted feature, i.e. all sequences that pass phase I selection must satisfy all filters in this phase. Most of the selective filters in Phase II are set to ensure the selection rule that the 5' antisense end should be less thermodynamically stable than the 5' sense end. This differential stability ensures that the antisense strand is incorporated into the RISC complex, reducing the unwanted off-target effect caused by the sense-strand [10,17,19,21,24,27,28,29]. In this study, the default cutoff for phase II selection is seven points, i.e. only those siRNA sequences that score seven points and above are considered functional. The scores of Phase III filters are reported for reference only. It would be useful for assessing the necessity of the existing and new rules. As part of the "Tuschl Rule [2]", many of the original siRNA selection software require the sense-strand to start with AA. However, this rule has been challenged recently because it filters out some potential effective siRNA sequences [17]. Therefore in this study, we set filter f-aa as optional.

### 2.3 BLAST and Smith-Waterman Search

Although the mechanism of siRNA's off-target effect is not fully understood, it is suggested that un-detected sequence homology by BLAST search may play a major role [17,34]. In this study, we employed two filters to screen for the possible off-target effect. First, BLAST is applied to identify and remove any off-target matches for all the siRNA sequences that survive the three-phase selection procedure. Then, the remaining sequences are screened by the Smith-Waterman search. By definition, both BLAST and Smith-Waterman are enabled and mandatory (much like the Phase I filters), but they are applied only to the sequences that passed all other filters.

### 2.4 The Implementation

The three-phase selection algorithm is implemented in Java so that it could be easily deployed as a web-based tool. The software accepts input of one or multiple target genes in Genbank or FASTA formats. Since the Genbank format provides locations for the coding region of the gene (CDS), it is the preferred format used in this study. Once the start location is determined for each gene

sequence, the selection process starts by collecting siRNA candidates. It shifts one nucleotide each time along the sequence to exhaust all potential siRNA sequences and avoids any sequences that contain uncertain nucleotides other than A, T/U, G, or C because these regions may have single nucleotide polymorphism, or SNP. The selection process is diagrammed in Figure 1.

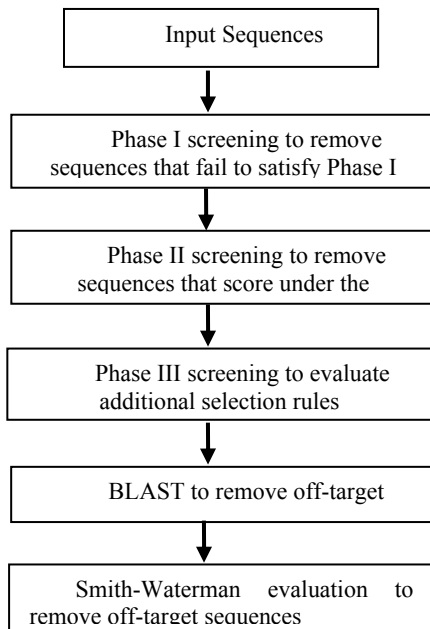


Figure 1. The flow chart of the siRNA selection process.

One of the major advantages of this tool is that it allows users to adjust all the selection criteria or even rearrange the filters in the three phases through a configuration file. Figure 2 shows an example where users can adjust the following from the graphic user interface (GUI) of this software: the length of the siRNA, the range of GC content and the definition of polymers of A, U/T, G and C, etc. The drop-down “Tool” menu shows other features of this software. The uses of both the BLAST and the Smith-Waterman searches are also selectable. However, whenever Smith-Waterman search is requested, BLAST is always performed first to minimize the computing time required for the Smith-Waterman search.

### 3 Results

To test the stringency of the default selection conditions described above, we applied them to the complete collection of human mRNAs in the NCBI RefSeq database (human genome build 35.1). This database contains 28,162 entries of which 27,956 are mRNA sequences, representing 23,484 protein-coding genes. Under such conditions, no suitable siRNA sequences could be found for 1915 genes (accounting for 2,075 entries, ~8.2% of the total genes). Further analysis reveals that the filters f-gc, f-poly, f-repeat and f-dgd are the major causes for those 1,915 genes to have zero siRNA sequence found. Of all the possible siRNA sequences from the 1,915 genes, 60.6% failed filter f-gc, 44.8% failed filter f-repeat, 76.4% failed filter f-poly and 65.9% failed filter f-dgd (while f-dgd is a selective filter, all others are mandatory in our default setting).

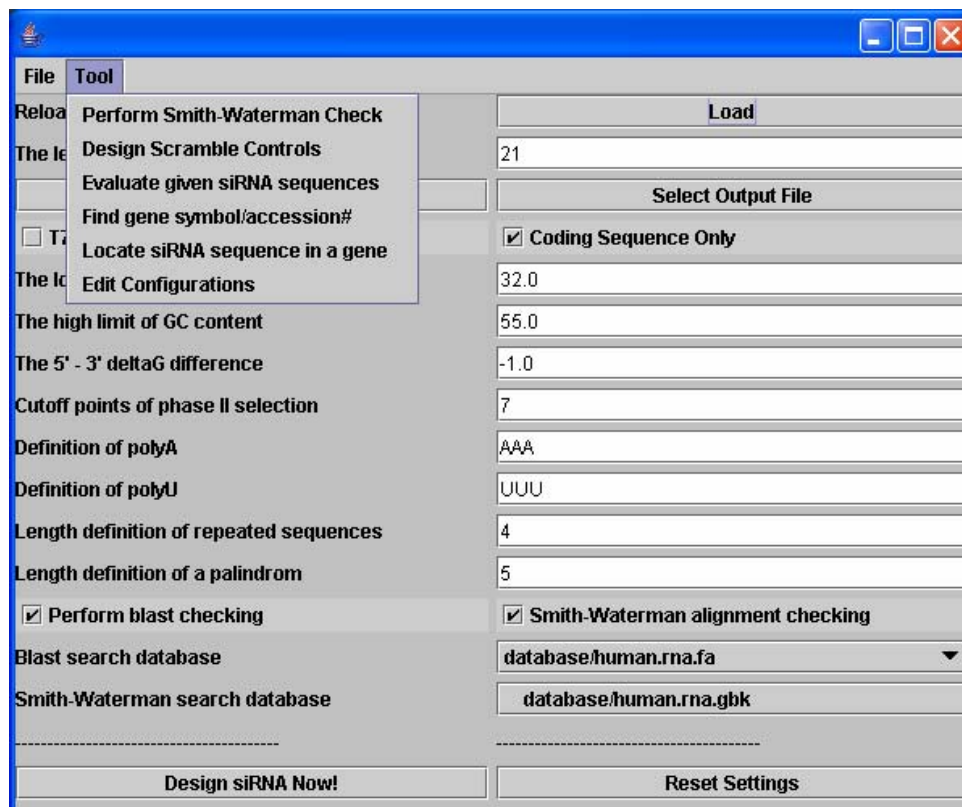


Figure 2. The graphic user interface (GUI) of the siRNA selection tool.

Interestingly, two among those 1,915 genes, PEN-2 (PSENE1, Genbank accession no. NM\_172341.1) and BIRC5 (Genbank accession no. NM\_001168.1) have functional siRNA sequences reported in the literature [36]. This result suggests that some modification of the rules has to be made in order to select the functional siRNA sequences for all genes.

In order to demonstrate the flexibility of the software, we modified the configuration file so that the definition for polymers (filter f-poly) is relaxed to accept AAAA, UUUU, GGGG and CCCC. With this single modification, the number of genes without a valid siRNA candidate reduced to 855 (from 1,915). Since some published siRNA sequences had GC content over 60%, we further modified the GC content limitation (filter f-gc) to be between 30 – 60%. Under this relatively less-stringent condition, the number of unsuccessful searches (855) is further reduced to 519, and valid siRNA sequences are found for the two genes PEN-2 and BIRC5 (although they are different from the published sequences). This experiment not only shows the flexibility of the three-phase algorithm, but also demonstrates its practicality of the whole package.

Another critical issue of siRNA design is to avoid any off-target effect. Although the true nature of off-target silencing of siRNA is yet to be elucidated, it has been suggested that the introduced siRNA will attack any mRNA sequences with less than 3 mismatches [17]. In order to demonstrate the ineffectiveness of using the BLAST filter alone in identifying those mismatches, we did the following experiments. As indicated in Table 1, we randomly chose 30 human genes and ran the three-phase selection program to get siRNA candidates before enabling the BLAST and Smith-Waterman filters. Then, about 100 siRNA candidates were randomly selected for BLAST and Smith-Waterman evaluation. After repeating this experiment 8 times, we found that about 66.6% of the siRNAs 19 bps in length could past the BLAST filter (minimum word size 7, gap penalty -1). However, after enabling the Smith-Waterman filter, we found that only 53.6% of those which passed BLAST test could survive the Smith-Waterman evaluation (gap penalty -3). Also shown in Table 1, the BLAST filter works better alone with longer siRNA sequences. For example, if the length of siRNA is set at 23 bps, it might be safe to assume the siRNA specificity without running the Smith-Waterman filter, because 99.7% of the BLAST-validated siRNAs could pass the Smith-Waterman evaluation.

To further validate our selection criteria, we collected 107 published siRNA sequences that targeted human genes. We found that only five of them could pass our default selection process. Close examination of the 102 failed sequences showed that 35 (34.3%) sequences failed the filter f-gc, 35 (34.3%) failed the filter f-repeat, 56 (54.9%) failed the filter f-poly and 68 (66.7%) failed the filter f-dgd. This result suggests that there could be many other better siRNA candidate sequences for these 107 published genes. A similar observation has been made by others [17].

	siRNA length (bps)		
	19	21	23
PB	66.6±4.0%	80.0±7.5%	87.4±6.9%
PSW	53.6±7.8%	98.6±1.6%	99.7±0.6%

**Table 1.** BLAST filter alone cannot safeguard the siRNA specificity. Experiments are repeated 8 times for about 100 randomly selected siRNA candidates generated from 30 randomly chosen gene sequences. Data is presented in the form of mean  $\pm$  standard deviation. PB: the percentage of siRNA candidates that can pass Blast test. PSW: the percentage of siRNA candidates that can pass Smith-Waterman test after passing Blast test.

Then we ran the 107 siRNA sequences through Smith-Waterman alignment with mismatch tolerance of 3 (where an insertion or a deletion accounts for 3 mismatches [24]). We have found that 32 sequences (representing 30 genes) failed this test. This indicates that some of the publicly validated siRNA sequences (as shown in Table 2) may risk off-target effect.

## 4 Discussion

The three-phase algorithm categorizes the major published siRNA design rules into three groups and applies them differentially in the design process based on their impacts on the siRNA function. Since all the rules are extracted from studying one or few genes, and there is little mechanistic justification for many of the rules, we should not treat those rules as absolute dogma. Rather, we should use those rules as a general guidance. The tool described in this paper provides the maximum flexibility for the user to adjust. Over time provided with sufficient experimental data input, this siRNA selection tool can be fine-tuned to provide intelligent design of highly effective siRNA on the whole-genome scale.

## Acknowledgement

The authors thank Dr. George J. Quellhorst, Jr. for critical reading of the manuscript.

## 5 References

- [1] A. Fire; S. Xu; M.K. Montgomery; S.A. Kostas; S.E. Driver; and C.C. Mello. "Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*." *Nature*, 391: 806-811, Feb 19 1998.
- [2] S.M. Elbashir; J. Harborth; W. Lendeckel; A. Yalcin; K. Weber; and T. Tuschl. "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells." *Nature*, 411: 494-498, 2001/05/24/print 2001.
- [3] P.J. Paddison; A.A. Caudy; and G.J. Hannon. "Stable suppression of gene expression by RNAi in mammalian cells." *PNAS*, 99: 1443-1448, February 5, 2002 2002.

Source	Target	Target-Symbol	Length	Sequence	Off-target matches
[37]	NM_000314	PTEN	19	CAAAUCCAGAGGCUAGCAG	NM_015245.1, 496-478, 2
[38]	NM_005163	AKT1	19	CCGCCAUCCAGACUGUGGC	XM_379163.1, 505-487, 2
[38]	NM_000321	RB1	19	GAUACCAGAUAUGUCAGA	NM_000132.2, 1916-1934, 2
[39]	NM_001904	CTNNB1	19	AGCUGAUUAUGAUGGACAG	XM_376254.2, 2346-2364, 2
[40]	NM_001838	CCR7	19	GAGGCUCAAGACCAUGACC	NM_000025.1, 395-413, 2
[40]	NM_001251	CD68	19	GCAAUAGCACUGCCACCAG	XM_373349.2, 656-638, 2
					NM_020528.1, 484-502, 2
[40]	NM_004355	CD74	19	ACUGACAGUCACCUCACCAG	NM_018407.3, 625-607, 2
[41]	NM_002483	CEACAM6	19	CCGGACAGUCCAUGUAUA	NM_001712.2, 477-494, 1
					NM_001815.1, 459-476, 1
					NM_133325.1, 727-745, 2
					NM_018288.2, 733-751, 2
[40]	NM_003467	CXCR4	19	CUGGCAUUGUGGGCAAUGG	NM_033104.2, 1940-1958, 2
					XM_497933.1, 5315-5333, 2
					NM_014974.1, 4236-4217, 2
[40]	NM_021095	SLC5A6	19	UAUUGGUUCCUGGGCUGCU	NM_020919.2, 4626-4644, 2
[40]	NM_001066	TNFRSF1B	19	CAGAACC CGAUCUGCACCU	NM_000302.2, 1202-1220, 2
					NM_002077.2, 2883-2901, 2
[42]	NM_024072	DDX54	19	GAAGAAGUCUGGAGGCUUC	NM_002022.1, 577-559, 2
					NM_138342.2, 1245-1227, 2
[43]	NM_002048	GAS1	19	UGGCGCUGCUGCAGCUGCU	115 off-target matches
[44]	NM_015895	GMNN	19	CUGGCAGAAGUAGCAGAAC	NM_014865.2, 968-986, 2
					(5 other off-target matches)
[45]	NM_012154	EIF2C2	19	UGGACAUCCCCAAAUAUGA	NM_198581.1, 4109-4127, 2
					(7 other off-target matches)
[46]	NM_001945	DTR	19	UACAAGGACUUCUGCAUCC	NM_080829.1, 745-763, 2
[47]	NM_001430	EPAS1	19	GCGACAGCUGGAGUAUGAA	NM_006023.1, 267-285, 2
[48]	NM_000599	IGFBP5	19	GAAGCUGACCCAGUCCAAG	NM_052839.2, 501-519, 2
					NM_198057.1, 365-383, 2
					NM_194278.2, 3341-3359, 2
[49]	NM_001278	CHUK	19	GCAGGCUCUUUCAGGGACA	NM_020746.1, 1069-1051, 2
					NM_019107.1, 643-625, 2
[50]	NM_032726	PLCD4	19	GGAAGGAGAAGAAUUCGUA	NM_002182.2, 1451-1469, 2
[51]	NM_004156	PPP2CB	19	UGUCUGCGAAAGUAUGGGA	XM_371140.3, 650-668, 2
[52]	NM_003253	TIAM1	19	GCGAAGGAGCAGGUUUUCU	NM_014065.2, 133-115, 2
					NM_017919.1, 1236-1254, 2
[53]	NM_006044	HDAC6	19	CCAGCCAAACCUAGGUUAG	XM_042234.6, 1855-1837, 2
					(8 other off-target matches)
[38]	NM_005030	PLK1	19	GUGCUUCGAGAUCUCGGAC	XM_498286.1, 528-546, 1
[38]	NM_005030	PLK1	19	GGGCAAGAUUGUGCCUAAG	XM_498286.1, 570-588, 0
[54]	NM_005053	RAD23A	20	AAGAGCCAUCAGAGGAAUC	NM_021574.1, 2290-2271, 2
[55]	NM_001274	CHEK1	21	GAAGCAGUCGAGUGAAGAUU	NM_002945.2, 359-379, 2
[56]	NM_052850	GADD45GIP1	21	AAGAUGCACAGAUGAUUGUG	XM_377715.1, 125-105, 2
[57]	NM_001419	ELAVL1	21	GUUGAAUCUGCAAAACUUAUU	XM_498103.1, 54-35, 1
[38]	NM_005030	PLK1	23	AAGGGCGCUUUGCCAAGUGCUU	XM_498286.1, 511-533, 0
[58]	NM_005573	LMNB1	23	AAGCUGCAGAUAGCUGGGCAA	NM_006258.1, 179-201, 2
[59]	NM_003302	TRIP6	23	AAGCCUACCACCCUGGCUGCUU	XM_059037.6, 1417-1439, 2

**Table 2.** Published siRNA sequences that may have off-target activities. Only the sense strand of the siRNA sequences are displayed. Off-target matches are arranged in order of gene accession number, the match position and the number of mismatches. If the start match position is larger than the stop match position, the homology is with the antisense strand of the searched gene.

- [4] J. Couzin. "BREAKTHROUGH OF THE YEAR: Small RNAs Make Big Splash." *Science*, 298: 2296-2297, December 20, 2002 2002.
- [5] M.L. Stephenson, and P.C. Zamecnik. "Inhibition of Rous sarcoma viral RNA translation by a specific oligodeoxyribonucleotide." *Proc Natl Acad Sci U S A*, 75: 285-288, Jan 1978.
- [6] L.J. Scherer, and J.J. Rossi. "Approaches for the sequence-specific knockdown of mRNA." *Nat Biotechnol*, 21: 1457-1465, Dec 2003.
- [7] S.M. Hammond; E. Bernstein; D. Beach; and G.J. Hannon. "An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells." *Nature*, 404: 293-296, Mar 16 2000.

- [8] E. Bernstein; A.A. Caudy; S.M. Hammond; and G.J. Hannon. "Role for a bidentate ribonuclease in the initiation step of RNA interference." *Nature*, 409: 363-366, Jan 18 2001.
- [9] G.J. Hannon. "RNA interference." *Nature*, 418: 244-251, 2002/07/11/print 2002.
- [10] A. Reynolds; D. Leake; Q. Boese; S. Scaringe; W.S. Marshall; and A. Khvorova. "Rational siRNA design for RNA interference." *Nat Biotechnol*, 22: 326-330, Mar 2004.
- [11] P.D. Zamore; T. Tuschl; P.A. Sharp; and D.P. Bartel. "RNAi: Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA at 21 to 23 Nucleotide Intervals." *Cell*, 101: 25-33, 2000/3/31 2000.
- [12] B.L. Bass. "RNA interference. The short answer." *Nature*, 411: 428-429, May 24 2001.
- [13] D.H. Kim; M. Longo; Y. Han; P. Lundberg; E. Cantin; and J.J. Rossi. "Interferon induction by siRNAs and ssRNAs synthesized by phage polymerase." *Nat Biotechnol*, 22: 321-325, Mar 2004.
- [14] 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, Feb 2002.
- [15] 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, Apr 15 2002.
- [16] C.C. Hardin; T. Watson; M. Corregan; and C. Bailey. "Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of d(CGCG3GCG)." *BIOCHEMISTRY*, 31: 833-841, 1992.
- [17] Y. Naito; T. Yamada; K. Ui-Tei; S. Morishita; and K. Saigo. "siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference." *Nucl. Acids Res.*, 32: W124-129, July 1, 2004 2004.
- [18] K. Ui-Tei; Y. Naito; F. Takahashi; T. Haraguchi; H. Ohki-Hamazaki; A. Juni; R. Ueda; and K. Saigo. "Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference." *Nucleic Acids Res*, 32: 936-948, 2004.
- [19] B. Yuan; R. Latek; M. Hossbach; T. Tuschl; and F. Lewitter. "siRNA Selection Server: an automated siRNA oligonucleotide prediction server." *Nucl. Acids Res.*, 32: W130-134, July 1, 2004 2004.
- [20] J. Martinez; A. Patkaniowska; H. Urlaub; R. Luhrmann; and T. Tuschl. "Single-stranded antisense siRNAs guide target RNA cleavage in RNAi." *Cell*, 110: 563-574, Sep 6 2002.
- [21] A. Khvorova; A. Reynolds; and S.D. Jayasena. "Functional siRNAs and miRNAs exhibit strand bias." *Cell*, 115: 209-216, Oct 17 2003.
- [22] E.P. Geiduschek, and G.A. Kassavetis. "The RNA polymerase III transcription apparatus1." *Journal of Molecular Biology*, 310: 1-26, 2001/6/29 2001.
- [23] G. Laughlan; A.I.H. Murchie; D.G. Norman; M.H. Moore; P.C.E. Moody; D.M.J. Lilley; and B. Luisi. "The high-resolution crystal structure of a parallel-stranded guanine tetraplex." *Science*, 265: 520-524, 1994.
- [24] W. Cui; J. Ning; U.P. Naik; and M.K. Duncan. "OptiRNAi, an RNAi design tool." *Comput Methods Programs Biomed*, 75: 67-73, Jul 2004.
- [25] N. Levenkova; Q. Gu; and J.J. Rux. "Gene specific siRNA selector." *Bioinformatics*, 20: 430-432, Feb 12 2004.
- [26] A.C. Hsieh; R. Bo; J. Manola; F. Vazquez; O. Bare; A. Khvorova; S. Scaringe; and W.R. Sellers. "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.
- [27] P. Dudek, and D. Picard. "TROD: T7 RNAi Oligo Designer." *Nucl. Acids Res.*, 32: W121-123, July 1, 2004 2004.
- [28] A. Henschel; F. Buchholz; and B. Habermann. "DEQOR: a web-based tool for the design and quality control of siRNAs." *Nucleic Acids Res*, 32: W113-120, Jul 1 2004.
- [29] P. S[ae]trom, and J. Snove, Ola. "A comparison of siRNA efficacy predictors." *Biochemical and Biophysical Research Communications*, 321: 247-253, 2004/8/13 2004.
- [30] S.P. Persengiev; X. Zhu; and M.R. Green. "Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs)." *Rna*, 10: 12-18, Jan 2004.
- [31] A.L. Jackson; S.R. Bartz; J. Schelter; S.V. Kobayashi; J. Burchard; M. Mao; B. Li; G. Cavet; and P.S. Linsley. "Expression profiling reveals off-target gene regulation by RNAi." *Nat Biotechnol*, 21: 635-637, Jun 2003.
- [32] P.C. Scacheri; O. Rozenblatt-Rosen; N.J. Caplen; T.G. Wolfsberg; L. Umayam; J.C. Lee; C.M. Hughes; K.S. Shanmugam; A. Bhattacharjee; M. Meyerson; and F.S. Collins. "Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells." *PNAS*, 101: 1892-1897, February 17, 2004 2004.
- [33] S.F. Altschul; W. Gish; W. Miller; E.W. Myers; and D.J. Lipman. "Basic local alignment search tool." *J Mol Biol*, 215: 403-410, Oct 5 1990.
- [34] O. Snove, Jr., and T. Holen. "Many commonly used siRNAs risk off-target activity." *Biochem Biophys Res Commun*, 319: 256-263, Jun 18 2004.
- [35] T.F. Smith, and M.S. Waterman. "Identification of common molecular subsequences." *J Mol Biol*, 147: 195-197, Mar 25 1981.
- [36] W.J. Luo; H. Wang; H. Li; B.S. Kim; S. Shah; H.J. Lee; G. Thinakaran; T.W. Kim; G. Yu; and H. Xu. "PEN-2 and APH-1 coordinately regulate proteolytic processing of presenilin 1." *J Biol Chem*, 278: 7850-7854, Mar 7 2003.

- [37] T.A. Vickers; S. Koo; C.F. Bennett; S.T. Crooke; N.M. Dean; and B.F. Baker. "Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis." *J Biol Chem*, 278: 7108-7118, Feb 28 2003.
- [38] D. Semizarov; L. Frost; A. Sarthy; P. Kroeger; D.N. Halbert; and S.W. Fesik. "Specificity of short interfering RNA determined through gene expression signatures." *PNAS*, 100: 6347-6352, May 27, 2003 2003.
- [39] U.N. Verma; R.M. Surabhi; A. Schmaltieg; C. Becerra; and R.B. Gaynor. "Small interfering RNAs directed against beta-catenin inhibit the in vitro and in vivo growth of colon cancer cells." *Clin Cancer Res*, 9: 1291-1300, Apr 2003.
- [40] S.J. Dunn; I.H. Khan; U.A. Chan; R.L. Searce; C.L. Melara; A.M. Paul; V. Sharma; F.Y. Bih; T.A. Holzmayer; P.A. Luciw; and A. Abo. "Identification of cell surface targets for HIV-1 therapeutics using genetic screens." *Virology*, 321: 260-273, Apr 10 2004.
- [41] M.S. Duxbury; H. Ito; M.J. Zinner; S.W. Ashley; and E.E. Whang. "CEACAM6 gene silencing impairs anoikis resistance and in vivo metastatic ability of pancreatic adenocarcinoma cells." *Oncogene*, 23: 465-473, Jan 15 2004.
- [42] R.R. Rajendran; A.C. Nye; J. Frasor; R.D. Balsara; P.G. Martini; and B.S. Katzenellenbogen. "Regulation of nuclear receptor transcriptional activity by a novel DEAD box RNA helicase (DP97)." *J Biol Chem*, 278: 4628-4638, Feb 14 2003.
- [43] R. Spagnuolo; M. Corada; F. Orsenigo; L. Zanetta; U. Deuschle; P. Sandy; C. Schneider; C.J. Drake; F. Breviario; and E. Dejano. "Gas1 is induced by VE-cadherin and vascular endothelial growth factor and inhibits endothelial cell apoptosis." *Blood*, 103: 3005-3012, Apr 15 2004.
- [44] H. Nishitani; Z. Lygerou; and T. Nishimoto. "Proteolysis of DNA replication licensing factor Cdt1 in S-phase is performed independently of geminin through its N-terminal region." *J Biol Chem*, 279: 30807-30816, Jul 16 2004.
- [45] H. Thonberg; C.C. Scheele; C. Dahlgren; and C. Wahlestedt. "Characterization of RNA interference in rat PC12 cells: requirement of GERP95." *Biochem Biophys Res Commun*, 318: 927-934, Jun 11 2004.
- [46] A. Gschwind; S. Hart; O.M. Fischer; and A. Ullrich. "TACE cleavage of proamphiregulin regulates GPCR-induced proliferation and motility of cancer cells." *Embo J*, 22: 2411-2421, May 15 2003.
- [47] O. Aprelikova; G.V. Chandramouli; M. Wood; J.R. Vasselli; J. Riss; J.K. Maranchie; W.M. Linehan; and J.C. Barrett. "Regulation of HIF prolyl hydroxylases by hypoxia-inducible factors." *J Cell Biochem*, 92: 491-501, Jun 1 2004.
- [48] P. Yin; Q. Xu; and C. Duan. "Paradoxical actions of endogenous and exogenous insulin-like growth factor-binding protein-5 revealed by RNA interference analysis." *J Biol Chem*, 279: 32660-32666, Jul 30 2004.
- [49] C. Kanei-Ishii; J. Ninomiya-Tsuji; J. Tanikawa; T. Nomura; T. Ishitani; S. Kishida; K. Kokura; T. Kurahashi; E. Ichikawa-Iwata; Y. Kim; K. Matsumoto; and S. Ishii. "Wnt-1 signal induces phosphorylation and degradation of c-Myb protein via TAK1, HIPK2, and NLK." *Genes Dev*, 18: 816-829, Apr 1 2004.
- [50] D.W. Leung; C. Tompkins; J. Brewer; A. Ball; M. Coon; V. Morris; D. Waggoner; and J.W. Singer. "Phospholipase C delta-4 overexpression upregulates ErbB1/2 expression, Erk signaling pathway, and proliferation in MCF-7 cells." *Mol Cancer*, 3: 15, May 13 2004.
- [51] A.V. Pandey; S.H. Mellon; and W.L. Miller. "Protein phosphatase 2A and phosphoprotein SET regulate androgen production by P450c17." *J Biol Chem*, 278: 2837-2844, Jan 31 2003.
- [52] A. Malliri; S. van Es; S. Huvener; and J.G. Collard. "The Rac exchange factor Tiam1 is required for the establishment and maintenance of cadherin-based adhesions." *J Biol Chem*, 279: 30092-30098, Jul 16 2004.
- [53] D. Girdwood; D. Bumpass; O.A. Vaughan; A. Thain; L.A. Anderson; A.W. Snowden; E. Garcia-Wilson; N.D. Perkins; and R.T. Hay. "P300 transcriptional repression is mediated by SUMO modification." *Mol Cell*, 11: 1043-1054, Apr 2003.
- [54] C. Brignone; K.E. Bradley; A.F. Kisselev; and S.R. Grossman. "A post-ubiquitination role for MDM2 and hHR23A in the p53 degradation pathway." *Oncogene*, 23: 4121-4129, May 20 2004.
- [55] J. Ahn; M. Urist; and C. Prives. "Questioning the role of checkpoint kinase 2 in the p53 DNA damage response." *J Biol Chem*, 278: 20480-20489, Jun 6 2003.
- [56] H.K. Chung; Y.W. Yi; N.C. Jung; D. Kim; J.M. Suh; H. Kim; K.C. Park; J.H. Song; D.W. Kim; E.S. Hwang; S.H. Yoon; Y.S. Bae; J.M. Kim; I. Bae; and M. Shong. "CR6-interacting factor 1 interacts with Gadd45 family proteins and modulates the cell cycle." *J Biol Chem*, 278: 28079-28088, Jul 25 2003.
- [57] M. Kullmann; U. Gopfert; B. Siewe; and L. Hengst. "ELAV/Hu proteins inhibit p27 translation via an IRES element in the p27 5'UTR." *Genes Dev*, 16: 3087-3099, Dec 1 2002.
- [58] J. Harborth; S.M. Elbashir; K. Bechert; T. Tuschl; and K. Weber. "Identification of essential genes in cultured mammalian cells using small interfering RNAs." *J Cell Sci*, 114: 4557-4565, Dec 2001.
- [59] F. Sanz-Rodriguez; M. Guerrero-Esteo; L.M. Botella; D. Banville; C.P. Vary; and C. Bernabeu. "Endoglin regulates cytoskeletal organization through binding to ZRP-1, a member of the Lim family of proteins." *J Biol Chem*, 279: 32858-32868, Jul 30 2004.