W124–W129 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh442

siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference

Yuki Naito[1], Tomoyuki Yamada[3], Kumiko Ui-Tei[1,2], Shinichi Morishita[3] and Kaoru Saigo[1,*]

[1]Department of Biophysics and Biochemistry, Graduate School of Science and [2]Undergraduate Program for Bioinformatics and Systems Biology, School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan and [3]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

Received February 15, 2004; Revised and Accepted February 20, 2004

ABSTRACT

siDirect (http://design.RNAi.jp/) is a web-based online software system for computing highly effective small interfering RNA (siRNA) sequences with maximum target-specificity for mammalian RNA interference (RNAi). Highly effective siRNA sequences are selected using novel guidelines that were established through an extensive study of the relationship between siRNA sequences and RNAi activity. Our efficient software avoids off-target gene silencing to enumerate potential cross-hybridization candidates that the widely used BLAST search may overlook. The website accepts an arbitrary sequence as input and quickly returns siRNA candidates, providing a wide scope of applications in mammalian RNAi, including systematic functional genomics and therapeutic gene silencing.

INTRODUCTION

siRNA (small interfering RNA) 21–22 nt in length, the active agent in RNAi (RNA interference), serves as a guide for cognate mRNA degradation (1,2). In mammals, siRNA is expected to become a powerful tool, not only for the large-scale gene silencing essential for functional genomics, but also for therapeutic purposes, including antiviral treatments (3–5). siRNA-mediated RNAi appears to include two distinct siRNA-sequence-dependent steps: the formation of the RNA-induced silencing complex (RISC) and the recognition of target mRNA (3,6).

siRNA introduced into cells interacts with PIWI-family proteins to form RISC (7,8). Before, or at an early stage of, RISC formation, siRNA duplexes are unwound preferentially at the thermodynamically less stable end through the action of helicase. Accordingly, the two siRNA strands (sense and antisense strands) may not be incorporated into the RISC equally (9). An siRNA strand with a less stable 5′ end (i.e. a 5′ end situated in a thermodynamically less stable duplex end) is incorporated into the RISC more efficiently than the opposite strand (9–11). Consequently, in cells transfected with a highly effective siRNA, RISC with the antisense strand, may predominate. Recently, practical guidelines for selecting highly effective siRNA sequences for mammalian RNAi have been proposed (11). The rules indicate that highly effective RNAi occurs in mammalian cells and chick embryos by siRNA that satisfies the following four conditions at the same time: (i) A/U at the 5′ end of the antisense strand, (ii) G/C at the 5′ end of the sense strand, (iii) AU-richness in the 5′ terminal third of the antisense strand and (iv) the absence of any GC stretch over 9 bp in length (Figure 1). These guidelines appear closely related to the molecular mechanism of RISC assembly as described previously (11). The first three guidelines may guarantee that the 5′ end of the antisense strand of siRNA is situated at or near the thermodynamically less stable siRNA duplex end.

The difficulty in designing siRNA sequences is the 'off-target' silencing effects of an siRNA duplex. Currently, we do not know how incomplete base-pairing between the antisense siRNA strand and non-targeted mRNA sequences

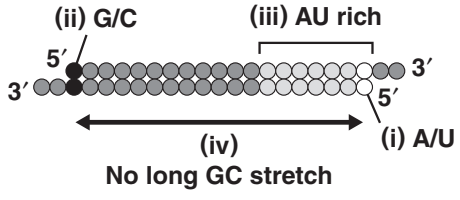

Figure 1. Structure of highly effective siRNA. See the text for details.

*To whom correspondence should be addressed. Tel: +81 3 5841 4407; Fax: +81 3 5841 4400; Email: saigo@biochem.s.u-tokyo.ac.jp
Correspondence may also be addressed to Shinichi Morishita. Tel: +81 47 136 3984; Fax: +81 47 136 3977; Email: moris@gi.k.u-tokyo.ac.jp

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

affects silencing. Recent studies have shown that even a single mismatch in the center of an siRNA can abolish silencing, indicating the remarkable specificity of siRNA-based RNAi that permits allele-specific gene silencing (12,13). In contrast, Jackson *et al*. (14) showed that as few as 11 contiguous matches between siRNA and an unrelated mRNA might cause off-target silencing. These results motivated us to design siRNA sequences that contain several mismatches to all non-targeted mRNA sequences.

This paper presents a new web-based online software system, siDirect, for selecting highly effective siRNA sequences with maximal target-specificity for mammalian RNAi. The siDirect algorithm incorporates the guidelines mentioned above to favor efficient mammalian RNAi with a high success rate, and it investigates all potential cross-hybridization candidates to avoid off-target gene silencing effects.

METHODS

Selecting highly effective siRNA sequences

Highly effective siRNA sequences are selected using an algorithm based on new guidelines developed by Ui-Tei *et al*. (11; Figure 1). Users can specify additional sequence conditions required for proper transcription initiation and termination (3,15) for designing short hairpin RNA, GC contents and custom rules.

The simple condition of requiring AA at the beginning of the target site is widely utilized to design siRNA sequences (2). This rule, however, frequently misses effective siRNA sequences that follow our four guidelines. Moreover, demanding AA at the beginning of the target site severely restricts the selection of possible siRNA sequence candidates that minimize off-target silencing effects. We therefore did not incorporate this rule into our system.

Reduction of off-target silencing effects

The minimization of off-target silencing effects calls for the selection of siRNA sequences that are guaranteed to have some mismatches to all unrelated sequences. We here use a rigorous specificity measure called the *mismatch tolerance*, the minimum number of mismatches between the siRNA sequence and any non-targeted sequence. For instance, an siRNA sequence of mismatch tolerance three does not match any off-target candidates with fewer than three mismatches. A higher mismatch tolerance of an siRNA sequence indicates its high specificity in the presence of some mismatches.

Exact mismatch-tolerance is costly to calculate, because it demands searching the entire sequence database to check whether individual siRNA oligos potentially cross-hybridize with irrelevant sequences. The Smith–Waterman local alignment algorithm (16) may return accurate answers but is very time-consuming to execute. In contrast, BLAST (17) is much faster than the Smith–Waterman algorithm, but it may overlook significant alignments.

BLAST may overlook off-target candidates

The following alignment illustrates such an example where BLAST fails to identify the similarity between two 19 nt sequences that match with three mismatches at the 5th, 10th and 14th positions, because the two sequences do not share seven contiguous base matches, which is the shortest word (consecutive nucleotides) that BLAST requires to find hits.

```
GAAGGCAGTCCAGTGAAAT  (NM_000014)
||||  ||||  |||  |||||
GAAGCCAGTACAGAGAAAT  (NM_002827)
```

Moreover, BLAST with its default parameter values may fail to notice best alignments with minimum number of mismatches when it receives such short sequences of 19 bases as input. For instance, using BLAST, we searched the 'nr' database for 'ACCGCAGTATATGGTTCTG', a 19 nt siRNA sequence candidate for NM_000014, and we received the partial answer that the first 15 nt matched a substring of NM_002864 at 100% identity. In fact, BLAST search overlooked the following best alignment with 18/19 matches.

```
ACCGCAGTATATGGTTCTG  (NM_000014)
|||||||||||||||  |||
ACCGCAGTATATGGTGCTG  (NM_002864)
```

This search failure is due to the default parameter values of BLAST. Since 'match reward' and 'mismatch penalty' are respectively set to 1 and −3, the occurrence of one mismatch demands at least four additional matches to extend the running alignment. A partial solution to fully extend such alignments is reducing the penalty of one mismatch, which, however, is likely to output numerous, low homologous alignments with off-target candidates.

To date, most existing websites for designing siRNA sequences, such as siRNA Target Finder at the Ambion website, siDESIGN Center at Dharmacon, siRNA Target Finder at GenScript (18), and Gene specific siRNA selector (19) use BLAST to search for off-target candidates. We should bear in mind that the limitations of BLAST in seeking optimal alignments may not minimize off-target silencing effects, as illustrated in the above example. In contrast, Qiagen utilizes SSearch, a rigorous Smith–Waterman search which is computationally costly. It was the lack of efficient, accurate software for enumerating potential off-target candidates that motivated us to develop an efficient method for computing mismatch tolerance (20).

Non-redundant sequence set of genes

Another major issue to solve was the generation of a non-redundant sequence set of genes for checking the target specificity of siRNA sequences. Traditional non-redundant sequence datasets, such as UniGene (21) and RefSeq (21), are not suitable for this purpose, because alternative splice variants in these datasets share common exons that bring about duplication. Although searching for siRNA oligos on common exons is valuable for simultaneous silencing of all alternative splice variants, one siRNA oligo may hybridize to any of the redundant exons, calling for duplicate elimination to yield one representative exon so that siRNA oligos can be properly designed. In addition, it is also necessary to consider siRNA sequences that target the junction connecting two

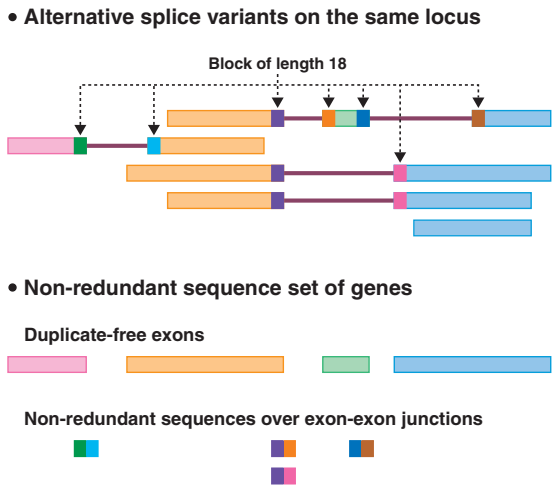

Figure 2. Generation of the non-redundant sequence set of genes from alternative splice variants located on the same locus.

exons of a particular alternative splice variant. Thus, non-redundant sequences over exon–exon junctions together with duplicate-free exons ought to comprise the non-redundant sequence set of genes for checking target specificity (see Figure 2).

Since such a database was not available, we created one. First we aligned all the human RefSeq and Unique UniGene sequences onto the human genomic sequences (hg16). For each query sequence, we selected the best alignment that had >90% coverage ratio and >85% match ratio. We retrieved duplicate-free exons and sequences over exon–exon junctions. However, since some sequences were not totally aligned to the genomic sequence, due to sequence errors or the incompleteness of the genomic sequence, subsequences that failed to match were added to the non-redundant sequence set. The non-redundant sequence set of mouse genes was similarly generated.

The major benefit of using the non-redundant sequence set is that the mismatch tolerance of a 'redundant sequence', defined as a substring of more than one exon on the same locus of the genome, is likely to be higher in the non-redundant sequence set than in the original set of human RefSeq and Unique UniGene sequences. We will present the statistics below.

Selection of target-specific siRNA sequences from the non-redundant sequence set

If an siRNA sequence is designed according to our four guidelines for effective sequences, the siRNA antisense strand is thought to be incorporated into the RISC more efficiently than the sense strand. This property may simply allow us to select effective siRNA sequences by considering only the sense target, the complement of the siRNA antisense strand, within the non-redundant sequence set. Thus, for siRNA sequences, we define, in particular, the *plus-strand mismatch tolerance* as mismatch tolerance calculated by using only the sense target sequence. However, we cannot disregard the possibility that the siRNA sense strand is also incorporated into the RISC and causes off-target effects. Thus, it is more reliable to take both strands of siRNA sequences into consideration. The *both-strand mismatch tolerance* is defined as the minimum number of mismatched bases that allow the siRNA antisense or sense strands to match a non-targeted sequence in the non-redundant sequence set. There remains the question of how much the mismatch tolerance of an siRNA sequence ought to be in order to treat the siRNA sequence target-specific.

The non-redundant sequence set of human genes was analyzed to obtain a comprehensive understanding of mismatch tolerance distribution of the 19 nt sequences that occur in the non-redundant set. The statistics for 19 nt sequences in Figure 3A shows that 9.5% are both-strand mismatch tolerance three or four but there exist no sequences of both-strand mismatch tolerance five or more. The fraction doubles if plus-strand mismatch tolerance is recalculated as illustrated in Figure 3B. From these results, we anticipate that effective siRNA sequences of both/plus-strand mismatch tolerance three or four can be designed for most of mRNA sequences, and we define a sequence to be *both-strand* (*plus-strand*, respectively) *specific* if the both-strand (plus-strand) mismatch tolerance is three or more. In reality, Figure 3C shows that, for 96.3% of mRNA sequences in RefSeq, at least one effective both-strand specific siRNA sequence is designed. The fraction increases to 97.7% if the plus-strand specificity is considered instead.

Figure 3D verifies the usefulness of the non-redundant sequence set, because in the original sequence set of human RefSeq and Unique UniGene, most of redundant 19 nt sequences are both-strand mismatch tolerance zero, while, in the non-redundant sequence set, 10.7% of redundant 19 nt sequences are both-strand specific and are therefore mismatch tolerance three or more.

Figure 4 illustrates the flowchart of siRNA sequence selection by our system. First, our web server accepts an arbitrary sequence or an accession number to retrieve its sequence. Subsequently, the query is processed to calculate effective, gene-specific siRNA sequences by searching the non-redundant sequence set for individual 19-nt sequences. To accelerate the computational performance, we precompute all the both/plus-strand specific siRNA sequences. This precomputation makes it possible to take just a couple of seconds to return the complete list of both/plus-strand specific siRNA sequences for a typical mRNA sequence (Figure 4B).

Care has to be taken to select an siRNA sequence since it may cross-react with off-target candidates. Our web server provides further information to examine the off-target silencing effects of a specific sequence. Clicking on the siRNA sequence asks the system to search the non-redundant sequence set for all the potential off-target candidates with which the siRNA sequence might cross-react. This complete search sounds to be computationally costly, but our algorithm (20) is capable of processing this request in less than a second. Subsequently, the server displays the alignment between each off-target candidate and the siRNA sequence in order to depict the locations of mismatches, which is useful in assessing the off-target silencing potential (Figure 4C).

Database maintenance

We plan to update our web server in response to major revisions to the human genome, the mouse genome, the RefSeq

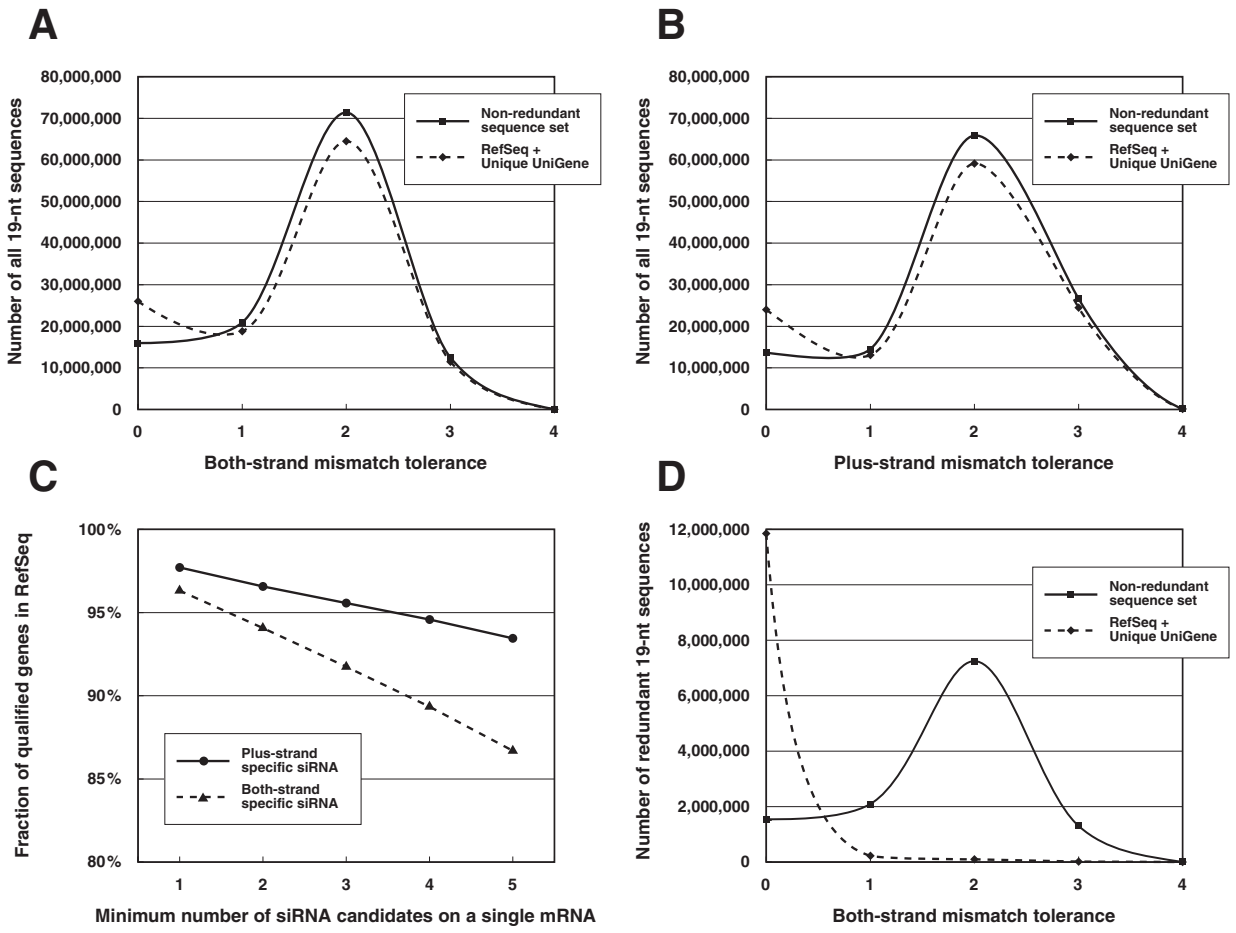

Figure 3. (**A**) The vertical axis is the number of 19 nt sequences of the both-strand mismatch tolerance shown in the horizontal axis. The solid line is the distribution for the non-redundant sequence set. For comparison, the dotted line shows the distribution when all the RefSeq and Unique UniGene sequences are used without removing any duplicates. Observe the dramatic reduction of redundant sequences of mismatch tolerance zero and the increases in the number of sequences of mismatch tolerance one or two. (**B**) The statistical chart (A) is recomputed for plus-strand mismatch tolerance. Note that the number of 19 nt sequences of mismatch tolerance three increases twofold. (**C**) The horizontal axis shows the requirement for the minimum number of effective, both/plus-strand specific siRNA candidates on a single mRNA. The vertical axis is the fraction of qualified genes in RefSeq that fulfill the constraint in the horizontal axis. Note that the fraction of qualified genes decreases severely when more both-strand specific siRNA sequences are designed, while the fraction decreases much more slowly when plus-strand specific siRNA are designed. This indicates the usefulness of the plus-strand specificity for designing multiple effective siRNA sequences on a single mRNA sequence. (**D**) The statistical chart in (A) is restricted to redundant 19 nt sequences and recalculated. This demonstrates that the non-redundant sequence set is indispensable for evaluating mismatch tolerances of redundant sequences correctly.

database and the UniGene database, though it is inevitable that such renewals will retract existing siRNA sequence candidates or add novel ones.

DISCUSSION

In reality, for ~2.3% of RefSeq sequences, we cannot design effective, plus-strand specific siRNA sequences. NM_010447, a mouse mRNA sequence, is a typical example. These cases are difficult to handle automatically. One may attempt to relax conditions on the target specificity by lowering the mismatch tolerance to two, which is likely to yield numerous effective siRNA sequence candidates, calling for some criteria for selecting more target-specific sequences from these numerous candidates.

Recall that Jackson *et al*. (14) stated that as few as 11 contiguous matches might cause off-target silencing. According to this observation, one promising measure would be the longest common factor of an siRNA sequence that is the length of the longest contiguous matches between the siRNA sequence and an unrelated sequence in the non-redundant set. A larger longest common factor is likely to indicate less possibility of cross-reaction with off-target candidates. Another criterion would examine the positions of mismatches. However, experimental confirmation of these criteria must be extensively performed to be utilized in practice, and hence we do not incorporate these measures into the current version of our website. Therefore, our recommendation is that users carefully select multiple siRNA sequences by investigating the list of off-target candidates (Figure 4C) and perform experiments to monitor off-target effects caused by individual siRNA sequences, even though the task is laborious.

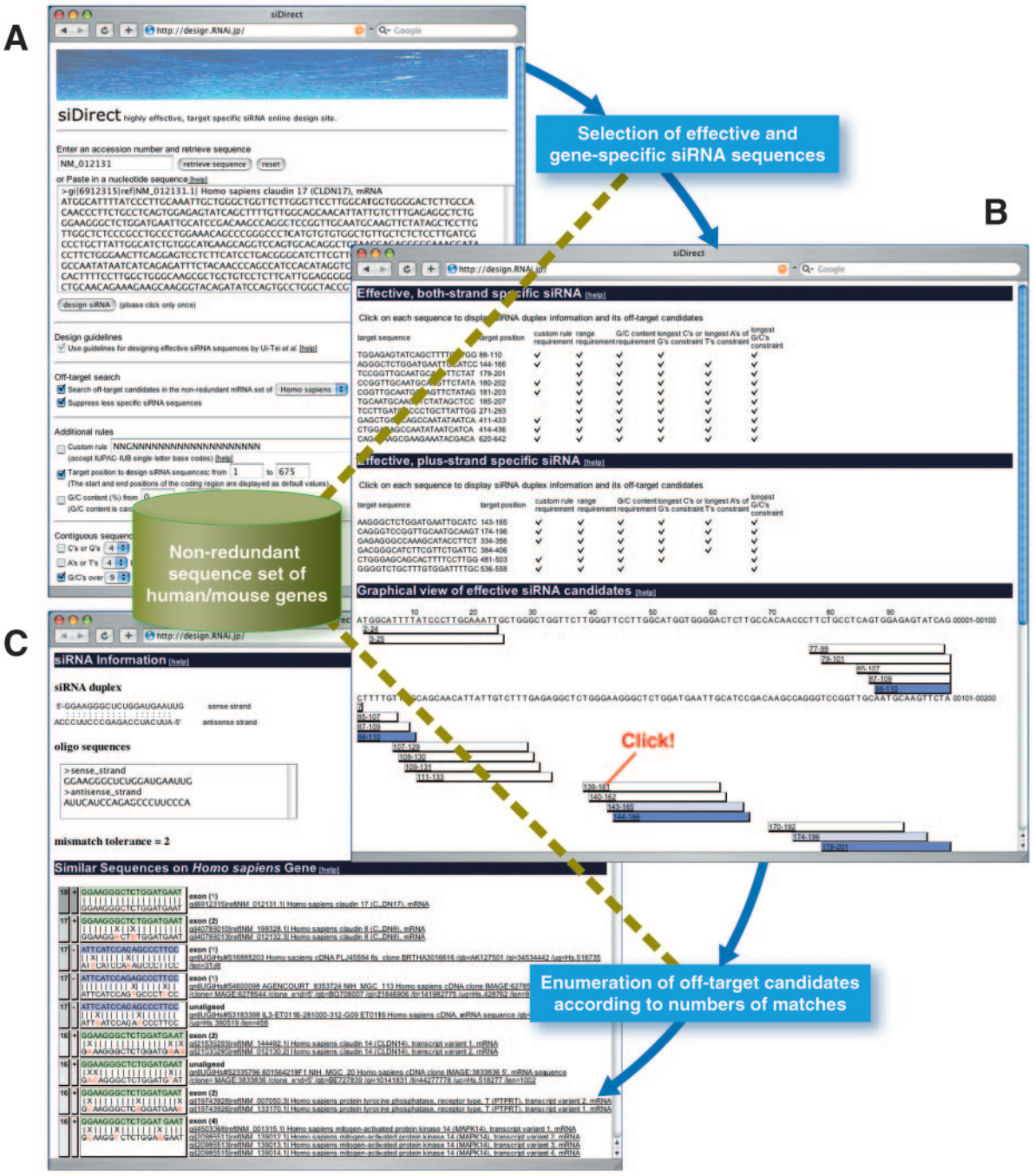

Figure 4. Flowchart of siRNA sequence selection by siDirect. (**A**) An arbitrary mRNA sequence is input. (**B**) Both/plus-strand specific precomputed siRNA sequences are presented in front. Both-strand specific (plus-strand specific, respectively) sequences are colored blue (light blue) and are placed under the mRNA sequence. Other siRNA sequences that meet the four guidelines of effective sequences are colored white. Clicking on an siRNA sequence displays the complete list of off-target candidates. (**C**) The alignment between each off-target candidate and the siRNA sequence clarifies the locations of mismatches.

ACKNOWLEDGEMENTS

We thank S. Zenno and F. Takahashi for helpful discussions and comments. The authors thank the anonymous referees for their suggestions, which have been valuable to improve our web server. This work was supported in part by funding from the Special Coordination Fund for Promoting Science and Technology to K.S., grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan to K.S. and K.U.-T. and a Grant-in-Aid for Scientific Research on Priority Areas (Grant #12208003) to S.M.

REFERENCES

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

2. Elbashir,S.M., Harborth,J., Lendeckel,W., Yalcin,A., Weber,K. and Tuschl,T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.
3. Dykxhoorn,D.M., Novina,C.D. and Sharp,P.A. (2003) Killing the messenger: short RNAs that silence gene expression. *Nature Rev. Mol. Cell Biol.*, **4**, 457–467.
4. Stevenson,M. (2003) Dissecting HIV-1 through RNA interference. *Nature Rev. Immunol.*, **3**, 851–858.
5. Gitlin,L. and Andino,R. (2003) Nucleic acid-based immune system: the antiviral potential of mammalian RNA silencing. *J. Virol.*, **77**, 7159–7165.
6. Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244–251.
7. Doi,N., Zenno,S., Ueda,R., Ohki-Hamazaki,H., Ui-Tei,K. and Saigo,K. (2003) Short-interfering-RNA-mediated gene silencing in mammalian cells requires Dicer and eIF2C translation initiation factors. *Curr. Biol.*, **13**, 41–46.
8. Martinez,J., Patkaniowska,A., Urlaub,H., Lührmann,R. and Tuschl,T. (2002) Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell*, **110**, 563–574.
9. Schwarz,D.S., Hutvágner,G., Du,T., Xu,Z., Aronin,N. and Zamore,P.D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.
10. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
11. Ui-Tei,K., Naito,Y., Takahashi,F., Haraguchi,T., Ohki-Hamazaki,H., Juni,A., Ueda,R. and Saigo,K. (2004) Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.*, **32**, 936–948.
12. Brummelkamp,T.R., Bernards,R. and Agami,R. (2002) Stable suppression of tumorigenicity by virus-mediated RNA interference. *Cancer Cell*, **2**, 243–247.
13. Miller,V.M., Xia,H., Marrs,G.L., Gouvion,C.M., Lee,G., Davidson,B.L. and Paulson,H.L. (2003) Allele-specific silencing of dominant disease genes. *Proc. Natl Acad. Sci. USA*, **100**, 7195–7200.
14. Jackson,A.L., Bartz,S.R., Schelter,J., Kobayashi,S.V., Burchard,J., Mao,M., Li,B., Cavet,G. and Linsley,P.S. (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnol.*, **21**, 635–637.
15. Shi,Y. (2003) Mammalian RNAi for the masses. *Trends Genet.*, **19**, 9–12.
16. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Wang,L. and Mu,F.Y. (2004) A web-based design center for vector-based siRNA and siRNA cassette, *Bioinformatics* (in press).
19. Levenkova,N., Gu,Q. and Rux,J.J. (2004) Gene specific siRNA selector. *Bioinformatics*, **20**, 430–432.
20. Yamada,T. and Morishita,S. (2004) Computing highly specific and noise-tolerant oligomers efficiently. *J. Bioinfo. Comp. Biol.* (in press).
21. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.