A Database-Based Two-Phase Algorithm for Efficient and Complete Detection of siRNA Off-Target Homology

Hong Zhou
Department of Mathematical Science, School of Health and Natural Sciences
University of Saint Joseph
1678 Asylum Avenue, West Hartford, CT 06117, USA
E-mail: hzhou@usj.edu

Hong Wang
Department of Chemical Biology and Center for Cancer Prevention Research
The State University of New Jersey
Rutgers, NJ 08854, USA
E-mail: howang@rci.rutgers.edu

Keywords: siRNA, shRNA, off-targets, RNAi, sequence alignment, Smith-Waterman

Received: June 31, 2014

Since the discovery of RNA Interference (RNAi), a cellular phenomenon in which a small double stranded RNA induces the degradation of its sequence specific target mRNA, using a computer-aided software tool to help design functional small interfering RNA (siRNA) or small hairpin RNA (shRNA) has become a standard procedure in applying RNAi to silence a target gene. A critical consideration in siRNA design is to avoid any possible off-target effect, i.e. to avoid sequence homology with untargeted genes. Though BLAST is the most powerful sequence alignment tool, it can overlook some significant homologies. Therefore, Smith-Waterman algorithm is the only approach that can guarantee to find all possible mismatch alignments that may cause off-target effect. However, Smith-Waterman alignment suffers from its inefficiency in searching through a large sequence database. A two-phase search algorithm was previously reported in which the first phase is used to identify local regions where the second phase, a bona fide Smith-Waterman alignment, is absolutely needed. Though such a two-phase homology search can improve the efficiency up to two orders of magnitude over the original Smith-Waterman alignment algorithm, it is still not efficient enough to be used alone for siRNA off-target homology search over a large sequence database. In this paper, we propose several improvements that dramatically speed up the reported two-phase algorithm while still guaranteeing the complete identification of siRNA off-target homologies.

Povzetek: V prispevku je predstavljena računalniška metoda za utišanje ciljnega gena.

1 Introduction

RNA interference (RNAi) is a cellular mechanism in which a small double stranded RNA induces the degradation of its sequence specific target mRNA, thus silencing the function of the target gene. Since its discovery, RNAi has become a powerful technique to knock out/down the expression of target genes for gene function studies in various organisms [3,5,16]. To employ this technique, the first step is to design target-specific small interference RNA (siRNA) or small hairpin RNA (shRNA) that is homologous to the target mRNA. Because of the predictability of RNAi based on its matching target sequence [2, 5, 7, 9–11, 14, 15, 19, 22, 25, 26], quite a few studies have been devoted to computer-guided algorithms to design effective siRNA or shRNA (from here on, this article will only refer to siRNA for simplicity) [4, 6, 12, 15, 20, 25, 26]. However, a critical requirement in siRNA design is to guarantee that the designed siRNA is free of off-target effect. Although the actual mechanism of off-target effect is still unknown, it has been demonstrated that a partial sequence homology between siRNA and its unintended targets is one of the major contributing factors [8,18,21]. It has been suggested that if an introduced siRNA has less than 3 mismatches with an unintended mRNA, it would likely knock down the expression of this mRNA in addition to its intended target which shares 100% sequence homology with this siRNA [11,15]. Unsurprisingly, the Basic Local Alignment Search Tool (BLAST) has been used to identify possible unintended homologous regions for siRNA candidates [1,13,17]. BLAST, although extremely fast, is not the best algorithm designed for this type of task since it overlooks significant sequence homologies [15,24,27]. As an alternative, Smith-Waterman alignment algorithm has been employed together with BLAST by

some design tools to identify all possible off-target sequences [15,27].

Smith-Waterman algorithm utilizes a dynamic programming approach to identify the local optimal alignment between two sequences [23]. It guarantees to locate the existing optimal alignment based on a scoring system with a set of scores assigned to a match, a substitution, a deletion, and an insertion. Given two sequences with length of *m* and *n*, the computational complexity of Smith-Waterman algorithm is $O(mn)$. Since the off-target search for siRNA sequences must be conducted completely through a given sequence database (which is usually large), the Smith-Waterman algorithm alone becomes very time-consuming and impractical for this task. Thus we once developed a two-phase homology search algorithm for siRNA off-target detection [29]. In this two-phase algorithm, the phase 1 procedure is used to identify the local regions where an off-target homology is possible to exist. Upon finding such local regions, the phase 2 procedure, a bona fide Smith-Waterman alignment algorithm, is used to determine if this local region has homology with the given siRNA sequence to cause off-target effect. This two-phase algorithm can be explained as the following.

For a siRNA of length *m*, an off-target homology is defined as a sequence that has less than *x* mismatches (i.e. mismatch cut-off equals *x*) when aligned against the siRNA (a mismatch is defined to be either a substitution, a deletion or an insertion hereafter). Thus, after the siRNA sequence is divided into *x* mutually disjointed and equal substrings (as equal as possible), at least one substring must have a perfect match with the off-target region. For the remainder of this paper, let's assume $m=21$ and $x=3$ unless stated otherwise. Under this condition, an off-target homology can only have a maximum of two mismatches, i.e., 0, 1, or 2 mismatches. When there are a maximum of two mismatches, no matter where the possible two mismatches are, at least one third of the siRNA sequence must have an exact match with the homological region. This concept is shown in Figure 1 which explains the case when the middle substring has the exact match.

Since all the possible off-target homological regions bear a substring of length 7 that has an exact match with the siRNA sequence, it is reasonable to perform the Smith-Waterman alignment only on the regions that have an exact match with at least one substring of the siRNA sequence. Thus, the first phase in the two-phase algorithm is designed to identify the potential regions with which at least one of the substrings of the siRNA sequence has an exact match. Only when such a potential region is identified, the second phase calls for the Smith-Waterman procedure to evaluate the best alignment between the potential region and the siRNA sequence. This algorithm does not construct any lookup table from the whole genome sequences, though it significantly improves the searching efficiency by guiding the most time-consuming core Smith-Waterman alignment on the local regions that need to be further examined.

Though the two-phase algorithm was shown to have efficiency gain of up to two orders of magnitude compared to the original Smith-Waterman algorithm alone [29], it is still not efficient enough to be applied alone for off-target homology search for a large number of siRNA sequences, such as the whole-genome siRNA design and off-target detection. For whole-genome siRNA design and off-target search, this two-phase algorithm must be applied with BLAST being the initial screening tool. In this paper, we present several significant improvements over both the phase 1 and the phase 2 procedures. These improvements dramatically speed up the original two-phase algorithm and make it able to complete off-target homology detection by itself alone for whole genome siRNA design.

2 Materials

The computer used in this study is a Dell notebook computer with Intel Core(TM) i5-2410M CPU. The maximum CPU speed is 2.30GHz. Installed RAM is 8.00 GB with 7.88 GB usable. The operating system is Windows 7 Enterprise (64 bit). The programming language used is Java.

The genome sequence database used in this study is NCBI human mRNA RefSeq gene database (human.rna.fna) downloaded on December 9, 2013. It has 68822 non-redundant sequences for mRNA/protein genes with average length of 3452 nucleotides.

The 1000 sample siRNA sequences used in this study were generated as the following: after 100 genes were randomly selected from the NCBI human mRNA RegSeq database, 10 siRNA were generated randomly from each gene using a computer-aided siRNA design tool [27]. All the siRNA sequences are of length 21 nucleotides (21-nt). One reason to select the length 21 is

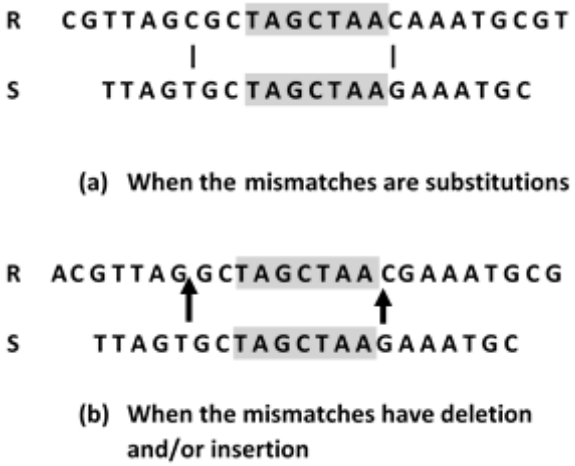

Figure 1: When there are 2 mismatches between the siRNA sequence (S) and the off-target region (R), at least one of the three substrings of S has the exact match with R. The vertical bars mark the substitutions, the arrows mark the deletion and/or insertion in R, and the shaded substrings have the exact match. When the substring in the middle of S has an exact match, the off-target region must be such a region in R that extends from the matched substring to both left and right enough base pairs to completely cover the siRNA

that 21-nt siRNA is the most commonly used siRNA in RNAi applications and the naturally occurring endo-siRNA is of 21-nt [30,31].

3 Improvements on phase 1

In the original two-phase algorithm, the phase 1 is nearly as five times time-consuming as the phase 2. This is shown in Table 1.

Table 1: The time cost (in seconds) analysis of the two-phase algorithm.

# of siRNA	Time Cost (seconds)	
	Phase-1	Phase-2
100	388.49	80.51
200	777.04	160.53
500	1938.97	404.02
1000	3866.47	798.10

The result in Table 1 is obtained by conducting off-target homology search through the whole human mRNA RefSeq gene database using the sample set of the 1000 21-nt siRNA. As the phase 1 is much more inefficient than the phase 2, our first improvement is on the phase 1.

The original two-phase algorithm used the Java's built-in string match algorithm, which is a character-by-charter brutal force algorithm. This algorithm has been shown to be inefficient in English text match. However, our experiment result shows that this brutal force algorithm performed equally efficient compared with both the Knuth-Morris-Pratt (KMP) algorithm and the Boyer-Moore algorithm in the siRNA off-target homology search. This is in fact unsurprising. There are only four different nucleotides in both DNA and RNA sequences, thus repeated sequences can occur frequently. The repeating sequences prohibit the skip-distance in both KMP and Boyer-Moore algorithms from growing, making them unable to achieve the desired efficiency gain.

The fact that there are only four different nucleotides in DNA sequences (let's use DNA as the example as the RefSeq database is for DNA) inspired us to develop a base-4 integer number system to represent DNA sequences. For example, let's define A=0, C=1, G=2 and T=3, then any nucleotide can be represented by a base-4 digit 0, 1, 2 or 3. Though the original two-phase algorithm works with siRNA of different lengths, in this study, siRNA of length 21 is used as the working sample. The reason is that 21-nt siRNA is the most commonly used and the naturally occurring endo-siRNA is of 21-nt [30,31]. In the NCBI probe database which contains thousands of siRNA sequences submitted by different researchers or companies, about 60% of these siRNA sequences are of length 21. However, please note that the concepts introduced in this study work for siRNA of different lengths. A 21-nt siRNA can be divided into three substrings each of size 7-nt. With the base-4 number system, any 7-nt can be represented by 7 digits, which is a base-10 integer between 0 and 16383 inclusively (please observe that $4^7 = 16384$). This means that a siRNA sequence can be represented by three base-4 integers each for a 7-nt subsequence. For example, a siRNA sequence of GCUGCAUCAACACAUGGAGCA is divided to three mutually disjointed 7-nt substrings GCUGCAU, CAACACA, UGGAGCA, which are represented as three integers 10131, 4164, and 14884 respectively. However, a DNA gene sequence of length M nucleotides must be represented by M-7+1 integers. This is because the homology search against the gene sequence is contiguous, shifting a nucleotide at a time. For example, AGCTATCCG is represented as an integer array of {2509, 10037, 7382}.

In the next experiment, we pre-processed the mRNA RefSeq database to convert every gene sequence into an array of integers. With this conversion, the phase 1 string match procedure becomes integer equivalence checking. It is not surprising to observe that the phase 1 procedure is significantly improved by representing the sequences as integers. The result is shown in Table 2.

Table 2: The time cost comparison between the original phase 1 and the modified phase 1 in which character by character comparison is transformed into integer comparison. (o): original Phase 1. (n): the new Phase 1 using integer comparison.

# of siRNA	Time Cost (seconds)		
	Phase1 (o)	Phase1 (n)	Phase2
100	388.49	164.20	80.51
200	777.04	345.81	160.53
500	1938.97	842.18	404.02
1000	3866.47	1760.54	798.10

Table 2 demonstrates that by using a base-4 integer system to represent the DNA nucleotides and thus transforming the string match process into an integer comparison process, the time cost of the original phase 1 can be cut down by more than 50%. The overall efficiency gain of the whole process is about 45%. Though the above experimental result is positive, the improvement is not significant enough. It is clear that dynamically searching for the exact match of a substring is always time-consuming. This motivated us to build a database to index the locations where each siRNA 7-nt substring has an exact match with the DNA gene sequences.

In the RefSeq database, there are 68822 non-redundant gene sequences with an average length of 3452 nucleotides. If we assume that all the four nucleotides have an equal chance to appear through the whole sequence database, then any a 7-nt subsequence has 1/16384 chance to appear, i.e. can show up about 14500 times in the whole gene database. To build the location-indexed database, we generated all the permutations (total 16384) of 7-nt, found the locations of each 7-nt in the RefSeq database and stored their location information in the location-indexed database. By using this location-indexed database, the phase 1 search process is no longer dynamic. Whenever a siRNA 7-nt substring needs to locate its exact matched regions inside the RefSeq gene database, using the integer

representation of the 7-nt substring as the primary key, such needed information is directly provided through this database. Via using such a database, the efficiency of the phase 1 is greatly improved. The result is shown in Table 3.

Table 3: By using a database to store the locations where each 7-nt substring has an exact match in the RefSeq gene sequence database, the phase 1 process is dramatically accelerated. Time-cost values are in seconds. (o): original Phase 1. (n): the new Phase 1 through the location-indexed database.

# of siRNA	Time Cost (seconds)		
	Phase-1 (o)	Phase-1 (n)	Phase-2
100	388.49	0.75	80.51
200	777.04	1.25	160.53
500	1938.97	3.09	404.06
1000	3866.47	5.96	798.10

Table 3 demonstrates that removing the dynamic searching process via a pre-built location-indexed database, the phase 1 process is speeded up by about 600 fold. Table 3 also shows that the phase 2 becomes now the bottleneck in the two-phase algorithm.

Because the phase 2 is now much slower than the phase 1 after using the pre-built database, the overall efficiency gain of the modified two-phase algorithm is only about 5 fold. The challenge becomes now, how to improve the phase 2.

4 Improvements on phase 2

The original phase 2 is a bona fide Smith-Waterman alignment algorithm. As the phase 1 is used to reduce the probability of using Smith-Waterman alignment in phase 2, we then tried to further reduce the use of the phase two operation by adding a pre-phase right before the original phase 2. This pre-phase serves as a filter to further remove unnecessary Smith-Waterman alignment.

The pre-phase dictates that only when the following two conditions are both met, Smith-Waterman alignment is needed.

Precondition: A 21-nt siRNA sequence (S) is equally divided into three mutually disjointed 7-nt substrings, S0, S1, and S3. When S0 finds an exact match with a substring R0 in region (R), the two other substrings of R would be R1 and R2, each corresponding to S1 and S2 separately.

Condition 1: Divide S1 from the middle to generate two sub-substrings. One is 3-nt, and the other is 4-nt. Repeat the dividing for S2. For each of the two corresponding substrings R1 and R2, extend one nucleotide to the direction away from R0 so that both R1 and R2 are of 8-nt. Check if R1 contains (the position does not need to match) either of the two sub-substrings of S1. Repeat the checking for R2. It must be true that the total number of sub-substrings contained in R1 and R2 is no less than 2.

Condition 2: Divide S1 as equally as possible to generate three mutually disjointed sub-substrings. One is 2-nt, one is 3-nt, and the last is 2-nt. Repeat the dividing with S2. For R1 and R2, extend two nucleotides to the direction away from R0 so that both R1 and R2 are of 9-nt. Check if R1 contains (the position does not need to match) any of the sub-substrings of S1, and repeat the checking for R2. It must be true that the total number of sub-substrings contained in R1 and R2 is no less than 4.

Figure 2 illustrates the condition 1.

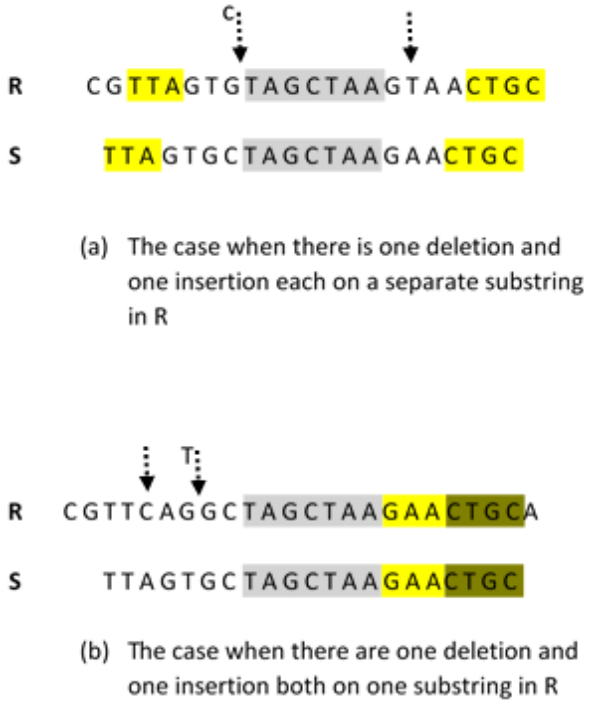

Figure 2: The Condition 1 in the case when the middle substring of siRNA finds an exact match in a region. In condition 1, there are always at least two sub-substrings of S that are contained inside R. Gray-shaded regions have the exact match. Arrows without a letter aside mark the insertion in R, and arrows with a letter aside mark the deletion (the letter indicates the deleted nucleotide in R). Yellow-shaded regions mark the sub-substrings of S contained in R.

The Condition 2 is depicted in Figure 3.

The first critical understanding of both Condition 1 and 2 is that when either S1 or S2 is divided into multiple sub-substrings, one mismatch, no matter what type it is, can only occur inside one sub-substring. Thus, in Condition 1, when there are four sub-substrings, at most two sub-substrings can be changed while at least two others are intact. Though a deletion or insertion can switch the positions of the sub-substrings, their content are not changed if the insertion/deletion are not inside the sub-substrings. A similar idea applies to Condition 2.

The second critical understanding of Condition 1 is that we need only to extend one nucleotide to the direction away from R0 so that both R1 and R2 are of 8-nt. The question raised here is that when R1 has two insertions, theoretically we need to extend two nucleotides so that R1 can fully cover S1. However, if R1 bears two insertions, given a homology between R and S, then S2 and R2 must be an exact match. Thus, there must be two sub-substrings of S2 that are contained

inside R2. It is then unnecessary to extend two nucleotides for R1 anymore. The similar idea can explain why it is necessary to extend only two nucleotides for both R1 and R2 in Condition 2.

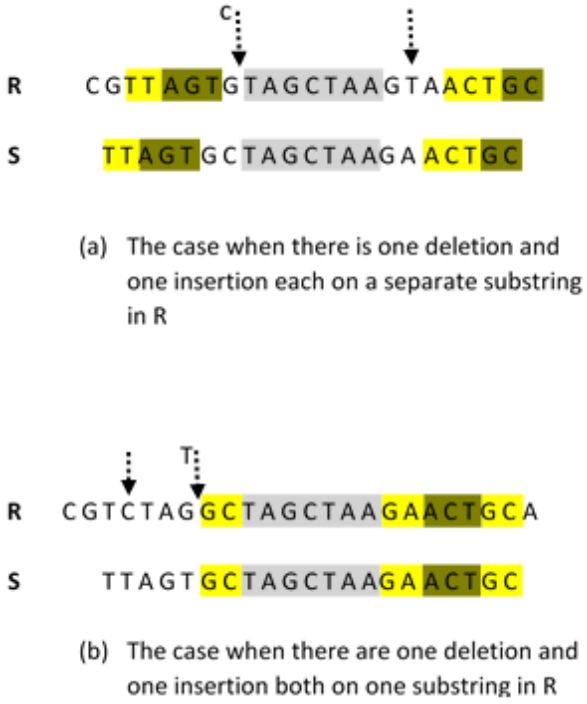

Figure 3: The Condition 2 in the case when the middle substring of siRNA finds an exact match in a region. In condition 2, there are always at least four sub-substrings of S that are contained inside R. Gray-shaded regions have the exact match. Arrows without a letter aside mark the insertion in R, and arrows with a letter aside mark the deletion (the letter indicates the deleted nucleotide in R). Yellow-shaded regions mark the sub-substrings of S contained in R.

With the pre-phase, the use of Smith-Waterman alignment is largely reduced and therefore the phase 2 is dramatically speeded up. The result is presented in Table 4.

Table 4: The pre-phase helps improve the efficiency of the phase 2 by more than 30 fold. (o) the old phase; (n) the new phase.

# of siRNA	Time Cost (seconds)		
	Phase1 (n)	Phase2 (o)	Phase2 (n)
100	0.75	80.51	2.60
200	1.25	160.53	49.46
500	3.09	404.06	12.68
1000	5.96	798.10	24.81

5 Discussion

The drawback of the phase 1 improvement is the necessity of building a database. Roughly speaking, for the 16384 different 7-nt substrings, there would be about 16384 x 14500 = 239018000 integers to store in the database, with each integer marking a position inside a gene for a 7-nt subsequence. In addition, there is other necessary information to store, such as the information of each gene. Depending on the implementation, the database size can be greater than or less than 1 Gb.

The pre-phase for phase 2 further reduces the use of Smith-Waterman alignment by mandating the satisfaction to both Condition 1 and Condition 2. Overall, the modified two-phase algorithm is 150 times more efficient than the original one. However, if only enforcing the satisfaction of one of the two conditions in the pre-phase, the improvement on efficiency is much less. By enforcing Condition 1 alone, the efficiency improvement on phase 2 is about 27 fold, while the efficiency improvement over the original phase 2 is only 11 fold if enforcing Condition 2 alone.

With the 1000 siRNA samples, there are 56402965 match hits in phase 1, indicating 56402965 alignment checking using Smith-Waterman algorithm in the original two-phase algorithm. However, there are only 399962 hits for the pre-phase. This shows that the pre-phase reduces the uses of Smith-Waterman alignment for about 140 fold. Among the 399962 hits, only 21444 of them were found to have true homology by Smith-Waterman alignment. This suggests that there might be additional approaches that can further improve the phase 2 efficiency.

Without considering the insertions or deletions, i.e. when only considering the case of substitutions, Smith-Waterman alignment is not necessary for the off-target homology detection. After the phase 1, for a homology with a maximum of two substitutions, the other two sub-strings in both siRNA and the searching region must have nearly exact matches with less than 3 substitutions. This is shown in Figure 4.

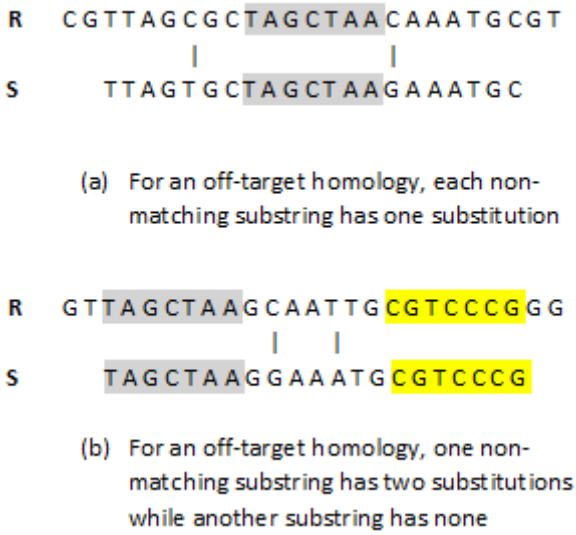

Figure 4: The case when only consider the substitutions in homology check. Gray-shaded regions indicate the exact match. Vertical bars mark the substitutions.

Since the substitutions do not change the positions of nucleotides, a check for the string matching on the two pairs of substrings can be quickly performed. The

experiment results show that it took only 4.00 seconds to complete the phase 2 for 1000 siRNA sequences. In addition, the experiment results disclose that there are only 21364 homologies found with only substitutions. Therefore, only 80 homologies identified for the 1000 siRNA sequences involve either deletions or insertions, a very small portion of the total number of off-target homologies (0.373%).

6 Conclusion

In the siRNA design, designing functional siRNA sequences is a relatively fast process, while the off-target evaluation is much more time consuming. Using the siRNA design tool [27], the time cost to design functional siRNA for all the 68822 human mRNA RefSeq non-redundant genes (an average of 33 siRNA for each gene) is about 400 seconds. With the improved two-phase algorithm (considering deletions and insertions), the time cost to completely check the off-target homology for all the designed siRNA sequences is estimated to be about 19.41 hours, which is acceptable for a process on the whole genome. Thus, after the improvements presented in this paper, the modified two-phase homology search algorithm can complete any off-target checking for functional siRNA design, without the initial use of BLAST.

Acknowledgement

The authors would like to thank Dr. Kevin Callahan and Dr. Joseph Manthey from the University of Saint Joseph for their critical reading of this manuscript.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., "Basic local alignment search tool," J Mol Biol., vol. 215, pp.403–410, 1990.

[2] Bass, B. L., "RNA interference. The short answer," Nature, vol.411, pp.428–429, 2001.

[3] Couzin, J., "BREAKTHROUGH OF THE YEAR: Small RNAs Make Big Splash," Science, vol.298, pp.2296–2297, 2002.

[4] Cui, W., Ning, J., Naik, U. P., Duncan, M. K., "OptiRNAi, an RNAi design tool," Comput Methods Programs Biomed., vol.75, pp.67–73, 2004.

[5] Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., Tuschl, T., "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells," Nature, vol.411, pp.494–498. 2001.

[6] Henschel, A., Buchholz, F., Habermann, B., "DEQOR: a web-based tool for the design and quality control of siRNAs," Nucleic Acids Res., vol.32, pp.W113–W120, 2004.

[7] Holen, T., Amarzguioui, M., Wiiger, M. T., Babaie, E., Prydz, H., "Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor," Nucleic Acids Res., vol.30, pp.1757–1766, 2002.

[8] Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G., "Expression profiling reveals off-target gene regulation by RNAi," Nat Biotechnol., vol.21, pp.635–637, 2003.

[9] Khvorova, A., Reynolds, A., Jayasena, S. D., "Functional siRNAs and miRNAs exhibit strand bias," Cell, vol.115, pp.209–216, 2003.

[10] Kim, D. H., Longo,M., Han, Y., Lundberg, P., Cantin, E., Rossi, J. J., "Interferon induction by siRNAs and ssRNAs synthesized by phage polymerase," Nat Biotechnol., vol.22, 321–325, 2004.

[11] Kim, D. H., Behlke,M. A., Rose, S. D., Chang, M. S., Choi, S., Rossi, J. J., "Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy," Nat Biotechnol., vol.23, pp.222–226, 2005.

[12] Levenkova, N., Gu, Q., Rux, J. J., "Gene specific siRNA selector," Bioinformatics, vol.20, pp.430–432, 2004.

[13] Lipman, D. J., Pearson, W. R., "Rapid and sensitive protein similarity searches," Science. vol.227, pp.1435–1441, 1985.

[14] Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., Tuschl, T., "Single-stranded antisense siRNAs guide target RNA cleavage in RNAi," Cell, vol.110, pp.563–574, 2002.

[15] Naito, Y., Yamada, T., Ui-Tei, K., Morishita, S., Saigo, K, "siDirect: highly effective, targetspecific siRNA design software for mammalian RNA interference," Nuclear Acids Research, vol.32, pp.W124–129, 2004.

[16] Paddison, P. J., Caudy, A. A., Hannon, G. J., "Stable suppression of gene expression by RNAi in mammalian cells," PNAS, vol.99, pp.1443–1448, 2002.

[17] Pearson, W. R., Lipman, D. J., "Improved tools for biological sequence comparison," PNAS, vol.85, pp.2444–2448, 1988.

[18] Persengiev, S. P., Zhu, X., Green, M. R., "Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs)," RNA, vol.10, pp.12–18, 2004.

[19] Reynolds, A., Leake, D., Boese, Q, Scaringe, S., Marshall, W. S., Khvorova, A., "Rational siRNA design for RNA interference," Nat Biotechnol., vol.22, pp.326–330, 2004.

[20] Sætrom, P., Snove, O. Jr., "A comparison of siRNA efficacy predictors," Biochemical and Biophysical Research Communications, vol.321, pp.247–253, 2004.

[21] Scacheri, P. C., Rozenblatt-Rosen, O., Caplen, N. J., Wolfsberg, T. G., Umayam, L., Lee, J. C., Hughes, C. M., Shanmugam, K. S., "Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells," PNAS, vol.101, pp.1892–1897, 2004.

[22] Siolas D., Lerner C., Burchard J, Ge W., Linsley P. S., Paddison P. J., "Synthetic shRNAs as potent

RNAi triggers," Nat Biotechnol., vol.23, pp.227–231, 2005.

[23] Smith, T. F.,Waterman,M. S., "Identification of common molecular subsequences," J Mol Biol., vol.147, pp.195–197, 1981.

[24] Snove, O., Jr., Holen, T., "Many commonly used siRNAs risk off-target activity," Biochem Biophys Res Commun., vol.319, pp.256–263, 2004.

[25] Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., Saigo, K., "Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference," Nucleic Acids Res., vol.32, pp.936–948, 2004.

[26] Yuan, B., Latek, R., Hossbach, M., Tuschl, T., Lewitter, F., "siRNA Selection Server: an automated siRNA oligonucleotide prediction server," Nucl. Acids Res., vol.32, pp.W130–W134, 2004.

[27] Zhou, H., Zeng, X., Wang, Y., Seyfarth, B. R., "A three-phase algorithm for computer aided siRNA design," Informatica (Slovene), 30 (2006) 357–364.

[28] Zhou, H., Zeng, X., Energy profile and secondary structure impact shRNA efficacy. BMC Genomics 2009, **10**(Suppl 1):S9.

[29] Zhou H, Wang Y, Zeng X: Fast and complete search of siRNA off-target sequences**.** In Proceedings of the international conference on bioinformatics & computational biology: 26–29 June 2006; Las Vegas. Edited by: Hamid R. Arabnia and Homayoun Valafar: CSREA Press; 2006:168-171.

[30] Carthew, R. W., Sontheimer, E. J., Origins and Mechanisms of miRNAs and siRNAs. Cell. 2009, 136(4): 642-655.

[31] Chen, L., Dahlstrom, J.E., Lee, S., Rangasamy, D., Naturally occurring endo-siRNA silences LINE-1 retrotransponsons in human cells through DNA methylation. Epigenetics, 7(7): 758-771, 2012.