A Practical Approach to Significance Assessment of siRNA Off-target Effects in RNA Interference

Wenzhong Zhao and Terran Lane
University of New Mexico, Department of Computer Science,
Albuquerque, NM 87131-0001, USA
{*wzhao, terran*}*@cs.unm.edu*

Abstract

Detection of potential cross-hybridization (or cross-reaction) between a short oligonucleotide sequence and a longer (unintended) sequence is crucial for many biological applications, such as selecting PCR primers, microarray nucleotide probes or short interfering RNAs (siRNAs). In this paper, we propose a flexible framework for estimating the significance of siRNA off-target effects on untargeted transcripts (messager RNA, or mRNA) in the RNA interference (RNAi) process. The framework can also be extended to other applications with minor changes.

We have developed and implemented a new homology sequence search framework – *siRNA Off-target Search* (SOS). SOS uses a hybrid, q-gram based approach, combining two filtering techniques using overlapping and non-overlapping q-grams. Our approach considers three types of imperfect matches based on biological experiments, namely G:U wobbles, mismatches, and bulges. The three main improvements over existing methods are: the introduction of a more general cost model (an affine bulge cost model) for siRNA-mRNA off-target alignment; the use of separate searches for alignments with and without bulges, that enables efficient discovery of potential off-target candidates in the filtration phase; and the use of position-preserving and order-preserving hit-processing techniques, that further improves the filtration efficiency. Overall, SOS achieves better performance, in terms of speed and recall/precision, than BLAST in detecting potential siRNA off-targets.

1 Introduction

RNA interference (RNAi) is a recently discovered posttranscriptional gene silencing (PTGS) mechanism that seems to play both regulatory and immunological roles in the eukaryotic genetic system [1, 9, 17, 23]. RNAi has aroused a great deal of excitement in both therapeutic and genomic experimental communities because of its potential for treatment of a wide spectrum of diseases such as HIV [11]; Huntington's diseases [25]; and certain classes of cancers [3, 9], in addition to its demonstrated use in functional genomic studies via controlled gene knockdown [5, 14].

At the heart of the RNAi cleavage event is the degree of matching between the target messager RNA (mRNA) and an initiator molecule, known as a short-interfering RNA (siRNA). By introducing a siRNA into a cell, we can induce the cellular machinery into degrading the mRNA product of a targeted gene and prevent further translation of the mRNA into protein. Thus, we can suppress the function of a specific (e.g., disease-related) gene. It was generally believed that RNAi process is highly specific [8, 7]. However, recent experimental results strongly suggest that siRNAs with imperfect matches can still knock down unintended mRNAs with high silencing efficacy [10, 19, 21]. Three types of imperfect matches have been studied in biological experiments: mismatches [10, 20], G:U wobbles [12, 20], and internal bulges [6]. In some cases, siRNAs can tolerate several mismatches to the target sequence [10]. A recent study [13] shows that about 75% of 359 published siRNAs have a risk of non-specific effects.

Designing highly effective and specific siRNAs is crucial for therapeutic or genomic applications of the RNAi process. siRNA efficacy has been studied extensively and design rules have been established for selecting effective siRNAs (e.g., [18, 23]). However, there is an urgent need to evaluate the significance of siRNA off-target reactions with unintended sequences. Since a siRNA recognizes its targets by sequence

complementarity, potential off-targets can be predicted by approximate sequence matching. However, this requires a pairwise sequence alignment between the siRNA and every gene in the genome, which can be very expensive for traditional sequence alignment algorithms such as dynamic programming.

Several programs that employ filtering techniques have been developed, including BLAST [2], PatternHunter [15], and QUASAR [4]. BLAST and PatternHunter filter out unrelated regions using contiguous and gapped seeds, respectively. They run much faster than dynamic programming, but both use lossy filtering techniques and thus frequently overlook off-target candidates [13]. QUASAR uses a q-gram based lossless filtering technique, but is limited to Hamming or Levenshtein distance alignments. There are also algorithms specific for siRNA selection (e.g., [13, 26]), but most deal only with mismatches.

We are interested in developing fast and lossless methods for evaluating the significance of siRNA off-target effects using approximate sequence matching. We define a flexible siRNA specificity measure called *off-target tolerance*: the maximum number of genes that have an *off-target score* less than or equal to a specified threshold with the selected siRNA. Our major contribution is the development and implementation of a new homology sequence search framework – *siRNA Off-target Search* (SOS) – which uses a hybrid, q-gram based approach, combining two filtering techniques using overlapping and non-overlapping q-grams. The three main improvements over existing methods are:

- the introduction of a more general cost model (an affine bulge cost model) for siRNA-mRNA off-target alignment;
- the use of separate searches for alignments with and without bulges, that enables more efficient discovery of potential off-target candidates; and
- the use of position-preserving and order-preserving hit-processing techniques, that further improves the filtration efficiency.

The rest of the paper is organized as follows. We introduce an affine bulge cost model as a measure for siRNA-mRNA off-target alignments in Section 2. In Section 3, we describe the two q-gram based filtering techniques for alignments without bulges and alignments with at least one bulge. We also discuss hit-processing techniques used to locate potential off-target candidates based on q-gram hit lists. We describe our computational experiments and report preliminary results in Section 4. Finally, we conclude and describe future work in Section 5.

2 siRNA-mRNA Distance-based Alignment

Consider a siRNA p, target gene g_i and mRNA $g_j \in G$, where G is the collection of genes in the genome. We define a *semi-global alignment* (alignment for short) between a siRNA and a mRNA to be a 5-tuple $A = \langle d, w, m, B_s, B_m \rangle$, where d, w, and m are the numbers of identical matches, G:U wobbles, and mismatches in the alignment; and $B_s = \{b_s\}$ and $B_m = \{b_m\}$ are the two sets of bulges on the siRNA and on the mRNA, respectively.

An affine bulge cost model is defined for computing the *alignment score*.

Definition 1 *Let $A = \langle d, w, m, B_s, B_m \rangle$ be an alignment. The affine* alignment score *for alignment A, $s(A)$, can be calculated as follows:*

$$s(A) = d\alpha + w\beta + m\gamma + \sum_{b_s \in B_s} (\rho + b_s\delta) + \sum_{b_m \in B_m} (\rho + b_m\delta),$$

where α, β, γ are the per-nucleotide scores for identity, G:U wobble, and mismatch; and ρ and δ are the scores for bulge creation and extension.

Let $N_s = \sum_{b_s \in B_s} b_s$ and $N_m = \sum_{b_m \in B_m} b_m$ be the total number of nucleotides in bulges on the siRNA and mRNA, respectively. We can rewrite the above formula as $s(A) = d\alpha + w\beta + m\gamma + (|B_s| + |B_m|)\rho + (N_s + N_m)\delta$.

We assume that $\alpha < \beta \leq \gamma$ and $0 \leq \delta \leq \rho$ hold for typical distance-based affine bulge cost models. A sample affine bulge cost model for siRNA off-target alignments is shown in Table 1. The numbers here are

Table 1: A sample affine bulge cost model for siRNA off-target alignments

Feature	Symbol	Score
Identity	α	0
G:U wobble	β	5
Mismatch	γ	10
Bulge creation	ρ	20
Bulge extension	δ	3

made up based on some experimental results of the effects of imperfect matches on RNAi process available in the literature [6, 10, 12, 20]. We use this cost model throughout the paper unless otherwise specified.

The typical length for a siRNA is 19-23 nucleotides long, while mRNAs are ∼2000 nucleotides long. Only a small portion of nucleotides in the mRNA contributes to an off-target alignment. The semi-global alignment for off-target detection does not penalize terminal bulges on the mRNA, but does penalize terminal bulges on the siRNA. Therefore, we define an *effective subsequence* of a mRNA in an alignment as follows:

Definition 2 *Given an alignment A between a siRNA and a mRNA, the* effective subsequence *of the mRNA in A is a portion of the contiguous nucleotides that aligns with the siRNA. The length of the effective subsequence is $L = N + N_m - N_s$, where N is the length of the siRNA.*

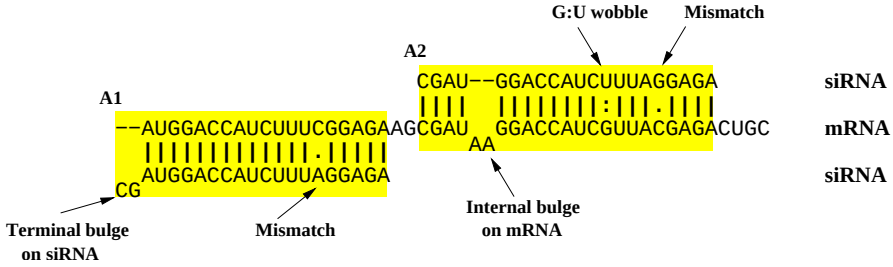

Figure 1: Semi-global alignments between a siRNA and a mRNA.

The following example is an illustration of semi-global off-target alignments between a siRNA and a mRNA.

Example 1 *Consider two off-target alignments between a siRNA and a mRNA, as shown in Figure 1. The off-target alignment score for alignment A_1: $s(A_1) = 18 \cdot 0 + 0 \cdot 5 + 1 \cdot 10 + (20 + 2 \cdot 3) + 0 = 36$, and for alignment A_2: $s(A_2) = 19 \cdot 0 + 1 \cdot 5 + 1 \cdot 10 + 0 + (20 + 2 \cdot 3) = 41$. The two effective subsequences of the mRNA in alignments A_1 and A_2 are shaded, and the effective lengths for the two alignments are 19 and 23, respectively.*

Let $\mathcal{A} = \{A | A$ is an alignment between a siRNA and a mRNA$\}$ be the set of all possible alignments between the siRNA and the mRNA. The *off-target score* between them, s, is defined to be the minimum alignment score, or $s = min_{A \in \mathcal{A}} s(A)$.

Finally, we define a siRNA specificity measure called *off-target tolerance*:

Definition 3 *Given a siRNA p, the target gene g_i and a genome $G = \{g_j | g_j$ is a gene in the genome.$\}$, the* off-target tolerance *is defined to be a tuple $\tau = \langle n, T \rangle$ such that there are at most n genes in G (except g_i) having off-target score $s \leq T$ with p.*

A siRNA sequence is considered gene-specific if there are at most n genes in the genome (other than the target gene) with off-target score $s \leq T$.

3 q-gram Based Filtering

Most fast sequence alignment algorithms use a two-phase approach: a filtration phase followed by a verification phase. A *filter* is a fast algorithm that discards large portions of the sequence according to some *filtering criterion*, leaving the remaining part to be checked in the verification phase.

Many filters in approximate sequence matching are based on *q-grams*: substrings of length q. Given a sequence S, a *positional q-gram* q_i of S is defined to be the substring of length q that starts at position i in the sequence; $i = 1, 2, \ldots, |S| - q + 1$. The basic idea behind q-gram based filtration is that the q-gram similarity between two sequences can be captured by the number of q-gram *hits* shared by them. A *hit* is a match between q-grams from the two sequences:

Definition 4 *A* hit $h(i, j)$ *between a siRNA p and a mRNA g is defined to be an exact, nucleotide-for-nucleotide match between q-gram q_i in p and q-gram q_j in g.*

Alignments between a siRNA and a mRNA may or may not have bulges. The score of a bulge in a typical affine bulge cost model is higher than that of a G:U wobble or a mismatch, so alignments with bulges require more identical matches than those without bulges in order to maintain the same alignment score. Therefore, the minimum number of q-gram hits that a potential off-target candidate must share with the siRNA is higher for alignments with bulges. By separating the searches for alignments with and without bulges, we raise the *filtering criterion* for searching alignments with bulges, thus reducing the number of potential off-target candidates to be checked in the verification phase. Based on this observation, we divide all alignments between a siRNA and a mRNA into two disjoint classes: one for alignments without bulges $\mathcal{A}_{\mathrm{w}}$, and the other for alignments with at least one bulge $\mathcal{A}_{\mathrm{b}}$, and $\mathcal{A} = \mathcal{A}_{\mathrm{w}} \cup \mathcal{A}_{\mathrm{b}}$.

The basic procedure for searching siRNA off-targets consists of the following four phases:

1. Lookup table creation phase: Build as an indexing structure a suffix array A over a sequence database G. Given the length q of q-grams, compute the start indexes of the hit lists for all q-grams in G. This step is performed once for G.

2. q-gram hit lists generation phase:

   a) Alignments without bulges: Generate non-overlapping q-grams for the selected siRNA, and search for the hit list for each q-gram.

   b) Alignments with at least one bulge: Generate overlapping q-grams for the selected siRNA, and search for the hit list for each q-gram.

3. Filtration phase:

   a) Use pigeonhole principle based approach (Corollary 1 in Section 3.1) along with hit-processing techniques (Section 3.3) to locate potential off-target candidates in G based on the q-gram hit lists from 2.a.

   b) Use q-gram lemma based approach (Corollary 2 in Section 3.2) along with hit-processing techniques (Section 3.3) to locate potential off-target candidates in G based on the q-gram hit lists from 2.b.

4. Verification phase:

   a) Potential off-target candidates from 3.a are further checked using direct sequence alignment with the siRNA.

   b) Potential off-target candidates from 3.b are further checked using traditional dynamic programming.

In the following sections, we first introduce two q-gram based approaches with overlapping and non-overlapping q-grams for determining the filtering criteria for potential off-target candidates. Then we describe the hit-processing techniques used to identify the potential off-target candidates based on the q-gram hit lists.

3.1 Filtering based on alignments without bulges

For simplicity, we represent an alignment without bulges as $A' = \langle d', w', m' \rangle$, and the alignment score $s(A') = d'\alpha + w'\beta + m'\gamma$.

To find the *filtering criterion* for potential off-target candidates based on alignments without bulges, we apply a non-overlapping q-gram based approach that is based on the pigeonhole principle lemma.

Lemma 1 *(Pigeonhole principle lemma [16]) Let s_1 and s_2 be two sequences of the same length l with Hamming distance k. If both s_1 and s_2 are divided into $\lfloor \frac{l}{q} \rfloor$ non-overlapping q-grams, then the number of q-gram hits between s_1 and s_2 is $t_w \geq \lfloor \frac{l}{q} \rfloor - k$.*

Here we extend the pigeonhole principle lemma to the new cost model for siRNA-mRNA off-target alignments without bulges.

Lemma 2 *Let $A' = \langle d', w', m' \rangle$ be an alignment between a siRNA and a mRNA. If both the siRNA and the effective subsequence of the mRNA (i.e., with length of N in this case) are divided into $\lfloor \frac{N}{q} \rfloor$ non-overlapping q-grams, then the number of q-gram hits between them is $t_w \geq \lfloor \frac{N}{q} \rfloor - (w' + m')$, where N is the length of the siRNA.*

With respect to q-grams, a G:U wobble is the same as a mismatch, so the total Hamming distance is just $(w' + m')$.

Given an off-target threshold, the following lemma gives an upper bound for the total number of G:U wobbles and mismatches $(w' + m')$.

Lemma 3 *Let $A' = \langle d', w', m' \rangle$ be an alignment between a siRNA and a mRNA with $s(A') \leq T$. The maximum number of G:U wobbles and mismatches in the alignment is $(m' + w') \leq \lfloor \frac{(1+\epsilon)(T-N\alpha)}{(\beta-\alpha)\epsilon+(\gamma-\alpha)} \rfloor$, where $\epsilon \geq 0$ is the ratio of the number of G:U wobbles to the number of mismatches.*[1]

Proof: $A' = \langle d', w', m' \rangle$ is an alignment without bulges, so the alignment score $s(A') = d'\alpha + w'\beta + m'\gamma$. Since $d' + w' + m' = N$, we have $d' = N - w' - m'$. Substituting d' and $w' = m'\epsilon$ into the alignment score, we have $s(A') = (N - m'\epsilon - m')\alpha + m'\epsilon\beta + m'\gamma$.

Rearranging the above equation yields $m' = \frac{s(A')-N\alpha}{(\beta-\alpha)\epsilon+(\gamma-\alpha)}$, so $m' + w' = (1+\epsilon)m' = \frac{(1+\epsilon)(s(A')-N\alpha)}{(\beta-\alpha)\epsilon+(\gamma-\alpha)}$. Since $s(A') \leq T$, we have $m'+w' \leq \frac{(1+\epsilon)(T-N\alpha)}{(\beta-\alpha)\epsilon+(\gamma-\alpha)}$. m' and w' are both integers, so $m'+w' \leq \lfloor \frac{(1+\epsilon)(T-N\alpha)}{(\beta-\alpha)\epsilon+(\gamma-\alpha)} \rfloor$.

Corollary 1 *Given an alignment $A' = \langle d', w', m' \rangle$ between a siRNA and a mRNA with $s(A') \leq T$, the number of non-overlapping q-gram hits between the siRNA and the effective subsequence of the mRNA is $t_w \geq \lfloor \frac{N}{q} \rfloor - \lfloor \frac{(1+\epsilon)(T-N\alpha)}{(\beta-\alpha)\epsilon+(\gamma-\alpha)} \rfloor$.*

Corollary 1 directly follows Lemmas 2 & 3, and can be used as a *filtering criterion* for determining potential off-target candidates based on alignments without bulges.

Example 2 *Consider an alignment $A' = \langle d', w', m' \rangle$ between a mRNA and a siRNA with length of $N = 21$ nucleotides. For a given off-target threshold T, the minimum number of q-gram hits t_w between the siRNA and the effective subsequence of the mRNA can be computed according to Corollary 1. The final results for the length of q-gram $q = 3$ are listed in Table 2.*

3.2 Filtering based on alignments with at least one bulge

To find the *filtering criterion* for potential off-target candidates based on alignments with bulges, we apply an overlapping q-gram based approach that is based on the q-gram lemma.

Lemma 4 *(The q-gram lemma[24]) Let p be a pattern and S be a target sequence with Levenshtein distance k. The number of overlapping q-gram hits between p and S is $t_b \geq |p| - (k+1)q + 1$.*

[1] In this paper, we assume a uniform distribution among the four nucleotides in genomes, therefore the average value for ϵ is around 0.2. Work on more general model of G:U wobble is ongoing.

Table 2: The minimum number of non-overlapping q-gram hits t_w, where $N = 21$, $q = 3$, and $\epsilon = 0.2$.

Off-target threshold T	q	t_w
0	3	7
10	3	6
20	3	5
30	3	4
40	3	3
50	3	2

Here we extend the *q-gram* lemma to the affine bulge cost model for siRNA-mRNA off-target alignments with bulges.

Lemma 5 *Given an alignment $A = \langle d, w,\ m, B_s, B_m\rangle$ between a siRNA and a mRNA, the number of overlapping q-gram hits between the siRNA and the effective subsequence of the mRNA is $t_b \geq N - q + 1 - (w + m)q - [|B_s|(q-1) + N_s] - |B_m|(q-1)$.*

The above formula can be split into four parts. The first part, $N - q + 1$, is the total number of q-grams (or valid q-gram hits) in the siRNA. The second part means that a single mismatch or G:U wobble, in the worst case, can invalidate up to q q-gram hits. The third and fourth parts represent the maximum numbers of q-grams that can be invalidated due to bulges on the siRNA and on the mRNA, respectively.

Given an off-target threshold, the following lemma gives an upper bound for the total number of G:U wobbles and mismatches $(w + m)$.

Lemma 6 *Let $A = \langle d, w, m, B_s, B_m\rangle$ be an alignment between a siRNA and a mRNA with $s(A) \leq T$. The maximum number of G:U wobbles and mismatches in the alignment is $(m + w) \leq \lfloor \frac{(1+\epsilon)[T-(N-N_s)\alpha-(|B|+|B'|)\rho-(N_s+N_m)\delta]}{(\beta-\alpha)\epsilon+(\gamma-\alpha)} \rfloor$, where $\epsilon \geq 0$ is the ratio of the number of G:U wobbles to the number of mismatches.*

Proof: $A = \langle d, w, m, B_s, B_m\rangle$ is an alignment, so the alignment score $s(A) = d\alpha + w\beta + m\gamma + (|B_s| + |B_m|)\rho + (N_s + N_m)\delta$. Since $d + w + m + N_s = N$, we have $d = N - w - m - N_s$. By substituting d and $w = m\epsilon$ into the alignment score, we get $s(A) = (N - m\epsilon - m - N_s)\alpha + m\epsilon\beta + m\gamma + (|B_s| + |B_m|)\rho + (N_s + N_m)\delta$.

Rearranging the above equation yields
$m = \frac{s(A)-(N-Ns)\alpha-(|B_s|+|B_m|)\rho+(N_s+N_m)\delta}{(\beta-\alpha)\epsilon+(\gamma-\alpha)}$, so
$m + w = (1 + \epsilon)m = \frac{(1+\epsilon)[s(A)-(N-Ns)\alpha-(|B_s|+|B_m|)\rho+(N_s+N_m)\delta]}{(\beta-\alpha)\epsilon+(\gamma-\alpha)}$. Since $s(A) \leq T$, we have
$m + w \leq \frac{(1+\epsilon)[T-(N-Ns)\alpha-(|B_s|+|B_m|)\rho+(N_s+N_m)\delta]}{(\beta-\alpha)\epsilon+(\gamma-\alpha)}$.
m and w are both integers, so
$m + w \leq \lfloor \frac{(1+\epsilon)[T-(N-Ns)\alpha-(|B_s|+|B_m|)\rho+(N_s+N_m)\delta]}{(\beta-\alpha)\epsilon+(\gamma-\alpha)} \rfloor$.

Corollary 2 *Given an alignment $A = \langle d, w, m, B_s, B_m\rangle$ between a siRNA and a mRNA with $s(A) \leq T$, the number of overlapping q-gram hits between the siRNA and the effective subsequence of the mRNA is $t_b \geq (N - N_s) - (|B_s| + |B_m| + 1)(q-1) - \lfloor \frac{(1+\epsilon)[T-(N-Ns)\alpha-(|B_s|+|B_m|)\rho+(N_s+N_m)\delta]}{(\beta-\alpha)\epsilon+(\gamma-\alpha)} \rfloor q$.*

Corollary 2 directly follows Lemmas 5 & 6, and can be used as a *filtering criterion* for determining potential off-target candidates based on alignments with at least one bulge.

Given an affine bulge cost model, an off-target threshold T, and the length q of q-grams, the minimum number of overlapping q-gram hits t_b between a siRNA and an effective subsequence of a mRNA depends on the following four parameters: $|B_s|, |B_m|, N_s$, and N_m. Therefore, we define $(t_b)_{min}$ to be the minimum number of overlapping q-gram hits t_b over all possible combinations of the four parameters. $(N_s)_{max}$ and $(N_m)_{max}$ are the maximum N_s and maximum N_m, respectively, such that $t_b > 0$.

Example 3 *Consider an alignment $A = \langle d, w, m, B_s, B_m\rangle$ between a mRNA and a siRNA with length of $N = 21$ nucleotides. For a given off-target threshold T, the minimum number of q-gram hits $(t_b)_{min}$ between the siRNA and the effective subsequence of the mRNA can be computed according to Corollary 2. The final results for the length of q-gram $q = 4$ are listed in Table 3, along with the $(N_s)_{max}$ and $(N_m)_{max}$.*

Table 3: The maximum N_s, maximum N_m, and the minimum number of overlapping q-gram hits $(t_b)_{min}$, where $N = 21$, $q = 4$, and $\epsilon = 0.2$ (Note that $(N_s)_{max}$ and $(N_m)_{max}$ need not be equal.).

Off-target threshold T	q	$(N_s)_{max}$	$(N_m)_{max}$	$(t_b)_{min}$
0, 10, 20	4	N/A	N/A	N/A
30	4	3	3	12
40	4	6	6	8
50	4	10	10	4

3.3 From q-gram hits to potential off-target candidates

Algorithm 1 Order-preserving Filtering

```
Input: A q-gram hit list H = {h(i, j)} in a sliding window, sorted in
       non-decreasing order of j
Output: Size of the maximum order-preserving hit list Ho ⊆ H
1  Order-preserving Filtering () {
     //Initialize size(i)
2    for each i=1 to |p|-q+1, do
3      size(i)=0;
     //Iteratively update size(i)
4    for each hit h(i, j) ∈ H, do
5      size(i)=max{size(i),max{size(k)|k<i}+1};
     //Return the maximum size
6    return max{size(i)};
7  }
```

For a fixed siRNA p, we first generate positional q-grams from p, $q_1, \ldots, q_n$. We use the lookup table to locate all q-gram hits in the genome for each q-gram q_i in p. Then we use hit-processing techniques to analyze the q-gram hit lists to determine potential off-target candidates. A good way of doing that is to use a sliding window of length W, and examine q-gram hits within the region of W contiguous nucleotides in the mRNA each time. A region is considered a potential off-target candidate if there are at least t (t_w for alignments without bulges or $(t_b)_{min}$ for alignments with bulges) q-gram hits within a sliding window. In order to guarantee that no potential off-target candidates will be overlooked, the length of a sliding window is at least as large as the maximum effective length of a mRNA for all possible alignments. Therefore, we define the length of a sliding window $W = N + (N_m)_{max}$.

Given a q-gram hit list, there are many ways to define potential off-target candidates with tradeoffs between filtration speed and efficiency. We describe three hit-processing techniques for defining potential off-target candidates. *Count filtering* simply counts the number of q-gram hits in a sliding window.

Proposition 1 Count filtering*: If A is an alignment between a siRNA p and a mRNA g with an alignment score $s(A) \leq T$, then there must exist a hit list of at least t q-gram hits between p and the effective subsequence of g.*

For alignments without bulges, the minimum number t in the proposition is t_w in Corollary 1. For alignments with bulges, the minimum number is $(t_b)_{min}$ in Corollary 2. In both cases, the minimum numbers remain the same for different hit-processing techniques.

While count filtering improves the efficiency of siRNA off-target search, it does not take advantage of ordering and positional information. Now, let's introduce the concept of an *order-preserving hit list.*

Definition 5 *A hit list H_o is an* order-preserving hit list *if and only if $j_l < j_k$ holds for any two hits $h(i_l, j_l), h(i_k, j_k) \in H_o$ and $i_l < i_k$.*

Proposition 2 Order-preserving filtering*: If A is an alignment between a siRNA p and a mRNA g with an alignment score $s(A) \leq T$, then there must exist an order-preserving hit list of at least t q-gram hits between p and the effective subsequence of g.*

Algorithm **1** uses the fact that q-gram hits in the same order-preserving hit list must appear in the same order as they appear in the siRNA. The time complexity for Algorithm **1** is $O(N|H|)$. Thus, the total time for the hit-processing is $O(N|\mathcal{H}|)$, where $\mathcal{H}$ is the joint hit list for the siRNA. Similarly, we define a *position-preserving hit list*.

Definition 6 *A hit list H_p is a* position-preserving hit list *if and only if* $|(j_k - i_k) - (j_l - i_l)| \leq (N_s)_{max} + (N_m)_{max}$ *for any two hits* $h(i_l, j_l), h(i_k, j_k) \in H_p$.

Algorithm 2 Position-preserving Filtering

```
Input: A q-gram hit list H = {h(i, j)} in a sliding window, sorted in
       non-decreasing order of (j − i)
Output: Size of the maximum position-preserving hit list Hp ⊆ H
1  Position-preserving Filtering () {
     //Initialization
2    span=(Ns)max+(Nm)max;
3    MAX_SIZE=0;
4    size=0;
5    start_diagonal=j_1-i_1;
6    for each hit h(i, j) ∈ H, do
7      current_diagonal=j-i;
8      if (current_diagonal - start_diagonal
           ≤ span), then
9        size++;
10     else
1        if (size>MAX_SIZE), then
2          MAX_SIZE=size;
         //Reset the start diagonal
13       size=1;
14       start_diagonal=current_diagonal;
     //Return the maximum size
15   return MAX_SIZE;
16 }
```

Proposition 3 Position-preserving filtering*: If A is an alignment between a siRNA p and a mRNA g with an alignment score $s(A) \leq T$, then there must exist a position-preserving hit list of at least t q-gram hits between p and the effective subsequence of g.*

Algorithm **2** uses the fact that q-gram hits in the same position-preserving hit list must be located on up to $(N_s)_{max} + (N_m)_{max}$ adjacent diagonals (i.e., A hit $h(i, j)$ is on the diagonal $(j - i)$.). The time complexity for Algorithm **2** is $O(|H|)$. Thus, the total time for the hit processing is $O(|\mathcal{H}|)$, where $\mathcal{H}$ is the joint hit list for the siRNA.

A more stringent filtering technique takes both the positional and ordering information. In order for a hit list of t hits to represent a potential off-target candidate, it must be an order-preserving and position-preserving hit list.

4 Computational Results

We have developed and implemented the *siRNA Off-target Search* (SOS) in Java. Here we report the results of our computational experiments. We first examine the performance of our SOS program, and then compare it with BLAST, a computer program commonly used for off-target detection. BLAST is downloaded from the NCBI ftp site. We performed all tests on a $3.0GHz$ Pentium IV machine running Linux with $1GB$ main memory. We apply our methods to the cDNA sequences of *C. elegans*, which contain 22,168 genes (database

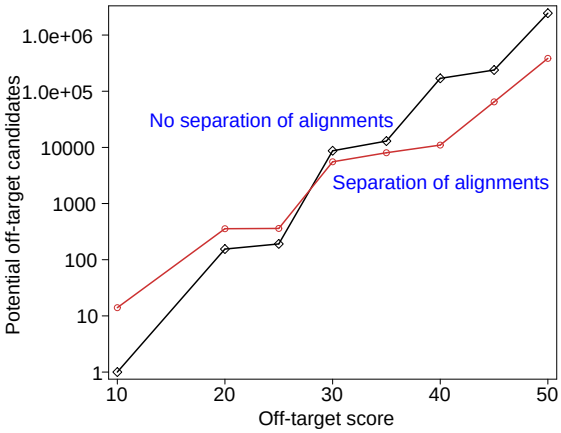

Figure 2: Effect of separate searches for alignments with and without bulges on the number of potential off-target candidates.

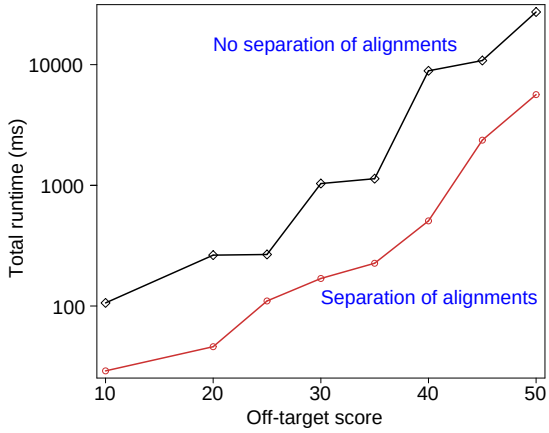

Figure 3: Effect of separate searches for alignments on running time.

size: 30MB) from release WS110 of the Wormbase at Sanger Institute [22]. Experiments were conducted to examine if separate searches for alignments with and without bulges, and the position-preserving and order-preserving hit-processing techniques improve the performance of the SOS. During the experiments, we collected both the number of potential off-target candidates after the filtration phase, which is an indicator of filtration efficiency, as well as the execution time. Each experiment was repeated with 100 randomly picked siRNAs, and each data point in the figures represents the average value of the results from those tests.

Figure 2 shows the effect of separate searches for alignments with and without bulges on the number of potential off-target candidates per siRNA. It can be seen that at lower off-target scores, separation of searches increases the potential off-target candidates, and the filtration efficiency decreases slightly. At higher off-target scores, separation of searches results in a ∼90% decrease of potential off-target candidates — the filtration efficiency increases dramatically.

Note that the number of potential off-target candidates is very low (less than 1000) at lower off-target scores, so the execution time in filtration phase dominates. However, the number of potential off-target candidates gets much higher at higher off-target scores, so the execution time in verification phase dominates. Therefore, the number of potential off-target candidates affects the overall performance only at higher off-target scores. This is supported by the fact that the total execution time with separate searches is consistently, for all off-target scores, almost one order of magnitude lower than that with no separation of searches, as shown in Figure 3.

The effects of hit-processing techniques on the number of potential off-target candidates and on running time are plotted in Figures 4 & 5. It is observed that both order-preserving and position-preserving hit-processing techniques reduce the number of potential off-target candidates in the filtration phase. The order-preserving hit-processing technique is especially effective in reducing the number of potential off-target candidates when searching for alignments with bulges (results not shown). Both techniques significantly reduce the total execution time, especially at higher off-target scores. The combination of both hit-processing techniques yields even better results than each individually.

We compared SOS with BLAST for siRNA off-target detection. SOS performs better than BLAST when matching a short sequence with a much longer sequence as in the siRNA off-target search problem. For a typical case (e.g., off-target threshold $T = 30$), SOS takes less than 0.2 second to finish the potential off-target search, as shown in Figure 5. Based on the execution time of 100 siRNA trials, BLAST takes an average of over 10 seconds for each siRNA with the default settings, which is at least one order of magnitude higher than that for SOS. Furthermore, BLAST missed a certain percentage of potential off-target sequences, as shown in Table 4. Similar recall/precision results have been seen for other genes as well. Both higher off-target threshold T and longer word length w contribute towards a higher rate of undetected off-targets.

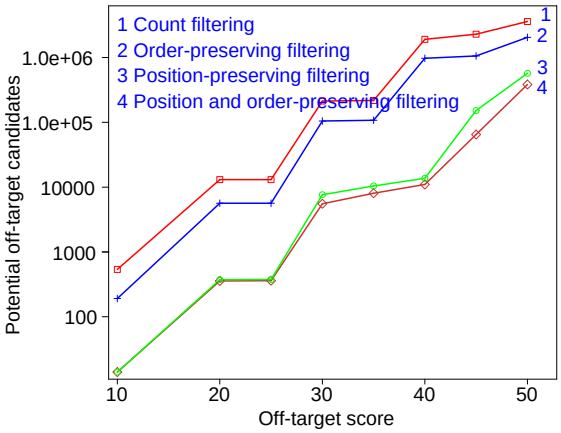

Figure 4: Effect of hit-processing techniques on the number of potential off-target candidates.

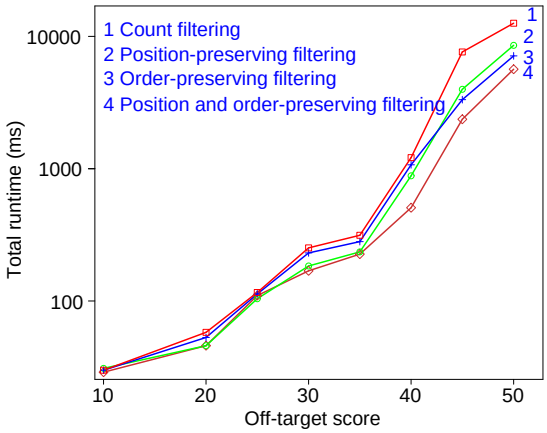

Figure 5: Effect of hit-processing techniques on running time.

Table 4: Percentage of potential siRNA off-targets not detected by BLAST given off-target threshold T and word length w.

Off-target threshold T	Total number of potential off-targets[a]	Percentage of off-targets not detected by BLAST[a]					
		$w = 6$	7	8	9	10	11
10	57	0.0	0.0	0.0	0.0	0.0	8.7
20	211	0.0	0.0	0.9	4.7	15.6	32.7
30	808	0.4	3.5	14.3	27.8	44.8	59.6
40	9906	3.7	15.6	32.9	50.7	65.9	76.6
50	300863	5.6	21.5	42.5	61.8	76.9	88.4

[a]These statistics are obtained based on tests for all siRNAs enumerated from a randomly picked gene (i.e., *B0024.1*, in this case) in the organism *C. elegans*.

5 Conclusions and Future Work

We have developed and implemented the *siRNA Off-target Search* (SOS) program. It uses a hybrid, q-gram based approach, combining two filtering techniques based on overlapping and non-overlapping q-grams. This approach introduces an affine bulge cost model to measure siRNA-mRNA off-target alignment. We have demonstrated with experiments that separate searches for alignments with and without bulges and the position-preserving and order-preserving hit-processing techniques significantly improve the filtration efficiency and running time. Overall, SOS achieves better performance, in terms of speed and recall/precision, than BLAST in detecting potential siRNA off-targets.

There are three major foci in our ongoing and future research: 1) Develop a specific method for G:U wobble detection in the filtration phase; 2) Use a more robust cost model considering positional information of imperfect matches; and 3) Apply gapped or partially matched q-grams in SOS.

Acknowledgement

This work was supported by NIH Grant Number 1P20RR18754 from the Institutional Development Award (IDeA) Program of the National Center for Research Resources.

References

[1] P. Ahlquist. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science*, 296:1270–1273, May 2002.

[2] S. F. Altschul, W. Gish, W. Miller, E. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[3] A. Borkhardt. Blocking oncogenes in malignant cells by RNA interference — new hope for a highly specific cancer treatment? *Cancer Cell*, 2(3):167–168, September 2002.

[4] S. Burkhardt, A. Crauser, H. P. Lenhof, E. Rivals, P. Ferragina, and M. Vingron. q-gram based database searching using a suffix array. In *Third Annual International Conference on Computational Molecular Biology*, pages 77–83, Lyon, 1999.

[5] J.-T. Chi, H. Y. Chang, N. N. Wang, D. S. Chang, N. Dunphy, and P. O. Brown. Genomewide view of gene silencing by small interfering RNAs. *PNAS*, 100(11):6343–6346, May 2003.

[6] Y. L. Chiu and T. M. Rana. RNAi in human cells: Basic structural and functional features of small interfering RNA. *Molecular Cell*, 10:549–561, 2004.

[7] A. Dillin. The specifics of small interfering RNA specificity. *Proc. the National Academy of Sciences (PNAS)*, 100(11):6289–6291, 2003.

[8] 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 c. elegans. *Nature*, 391:806–811, 1998.

[9] G. J. Hannon. RNA interference. *Nature*, 418:244–251, July 2002.

[10] 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. *Nature Biotechnology*, 21(6):635–638, 2003.

[11] J. M. Jacque, K. Triques, and M. Stevenson. Modulation of HIV-1 replication by RNA interference. *Nature*, 418:435–438, July 2002.

[12] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks. Human MicroRNA targets. *PLos Biology*, 2(11):1862–1879, 2004.

[13] O. Snove Jr. and T. Holen. Many commonly used siRNAs risk off-target activity. *Biochemical and Biophysical Research Communications*, 319:256–263, 2004.

[14] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, Sohrmann M., Welchman D. P., P. Zipperlen, and J. Ahringer. Systematic functional analysis of the caenorhabditis elegans genome using RNAi. *Nature*, 421:231–237, January 2003.

[15] B. Ma, J. Tromp, and M. Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.

[16] E. W. Myers. A sublinear algorithm for approximate keywords searching. *Algorithmica*, 12:345–374, 1994.

[17] R. H. A. Plasterk. RNA silencing: The genome's immune system. *Science*, pages 1263–1265, May 2002.

[18] A. Reynolds, D. Leake, Q. Boese, S. Scaring, W. Marshall, and A. Khvorova. Rational siRNA design for RNA interference. *Nature Biotechnology*, 22(3):326–330, 2004.

[19] X. Zhu S. P. Persengiev and M. R. Green. Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). *RNA*, 10(1):12–18, 2004.

[20] S. Saxena, Z. O. Jonsson, and A. Dutta. Small RNAs with imperfect match to endogenous mRNA repress translation: implications for off-target activity of siRNA in mammalian cells. *J. Biol. Chem.*, 278(45):44312–44319, 2003.

[21] 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. *Proc. Natl. Acad. Sci.*, 100(11):6347–6352, 2003.

[22] The Sanger Institute. From Wormbase - the C. elegans genome database. http://www.wormbase.org/, February, 2004.

[23] T. Tuschl. RNA interference and small interfering RNAs. *Chembiochem*, 2(4):239–245, 2001.

[24] E. Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92:191–211, 1992.

[25] H. Xia, Q. Mao, S. L. Eliason, S. Q. Harper, I. H. Martins, H. T. Orr, H. L. Paulson, L. Yang, R. M. Kotin, and B. L. Davidson. RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. *Nature Medicine*, 10:816–820, July 2004.

[26] T. Yamada and S. Morishita. Accelerated off-target search algorithm for siRNA. *Bioinformatics, Advance Access published on December 14, 2004*, 0(1552), 2004.