Published online March 30, 2005

1834–1847 *Nucleic Acids Research, 2005, Vol. 33, No. 6*
doi:10.1093/nar/gki324

A computational study of off-target effects of RNA interference

Shibin Qiu, Coen M. Adema[1] and Terran Lane*

Department of Computer Science and [1]Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA

Received December 13, 2004; Revised February 19, 2005; Accepted March 7, 2005

ABSTRACT

RNA interference (RNAi) is an intracellular mechanism for post-transcriptional gene silencing that is frequently used to study gene function. RNAi is initiated by short interfering RNA (siRNA) of ~21 nt in length, either generated from the double-stranded RNA (dsRNA) by using the enzyme Dicer or introduced experimentally. Following association with an RNAi silencing complex, siRNA targets mRNA transcripts that have sequence identity for destruction. A phenotype resulting from this knockdown of expression may inform about the function of the targeted gene. However, 'off-target effects' compromise the specificity of RNAi if sequence identity between siRNA and random mRNA transcripts causes RNAi to knockdown expression of non-targeted genes. The complete off-target effects must be investigated systematically on each gene in a genome by adjusting a group of parameters, which is too expensive to conduct experimentally and motivates a study *in silico*. This computational study examined the potential for off-target effects of RNAi, employing the genome and transcriptome sequence data of *Homo sapiens*, *Caenorhabditis elegans* and *Schizosaccharomyces pombe*. The chance for RNAi off-target effects proved considerable, ranging from 5 to 80% for each of the organisms, when using as parameter the exact identity between any possible siRNA sequences (arbitrary length ranging from 17 to 28 nt) derived from a dsRNA (range 100–400 nt) representing the coding sequences of target genes and all other siRNAs within the genome. Remarkably, high-sequence specificity and low probability for off-target reactivity were optimally balanced for siRNA of 21 nt, the length observed mostly *in vivo*. The chance for off-target RNAi increased (although not always significantly) with greater length of the initial dsRNA sequence, inclusion into the analysis of available untranslated region sequences and allowing for mismatches between siRNA and target sequences. siRNA sequences from within 100 nt of the 5′ termini of coding sequences had low chances for off-target reactivity. This may be owing to coding constraints for signal peptide-encoding regions of genes relative to regions that encode for mature proteins. Off-target distribution varied along the chromosomes of *C.elegans*, apparently owing to the use of more unique sequences in gene-dense regions. Finally, biological and thermodynamical descriptors of effective siRNA reduced the number of potential siRNAs compared with those identified by sequence identity alone, but off-target RNAi remained likely, with an off-target error rate of ~10%. These results also suggest a direction for future *in vivo* studies that could both help in calibrating true off-target rates in living organisms and also in contributing evidence toward the debate of whether siRNA efficacy is correlated with, or independent of, the target molecule. In summary, off-target effects present a real but not prohibitive concern that should be considered for RNAi experiments.

INTRODUCTION

RNA interference (RNAi) (1) is an intracellular mechanism for post-transcriptional gene silencing that most probably functions in the regulation of gene expression and defense against transposable DNA elements and viruses. RNAi is triggered by double-stranded RNA (dsRNA). Dicer, an enzyme with RNAse activity, cleaves dsRNA into fragments of ~21 nt, termed short interfering RNA (siRNA). The siRNA associates with several proteins to form an RNAi silencing complex (RISC). The sequence of the minus-strand of the siRNA then

*To whom correspondence should be addressed at Department of Computer Science, University of New Mexico, Farris Engineering Building Room 325, Albuquerque, NM 87131-1386, USA. Tel: +1 505 277 9609; Fax: +1 505 277 9627; Email: terran@cs.unm.edu

targets mRNA molecules that have sequence identity for cleavage by RISC. This sequence-directed removal of particular mRNA transcript yields a knockdown of expression of the affected gene. Extensive investigations are ongoing to gain more detailed understanding of RNAi. RNAi has been widely used as an experimental tool for the study of gene function and can be applied for large-scale analyses (2–4). RNAi has aroused a great deal of excitement in both therapeutic and genomic experimental communities because of its potentials for the treatment of a wide spectrum of diseases, such as HIV (5,6), spinocerebellar ataxia type 1 and Huntington's diseases (7), certain classes of cancers (8–10) and hypercholesterolemia (11,12), as well as its demonstrated use in functional genomic studies via controlled gene knockdown (13–15).

Both dsRNA and siRNA have been used to knockdown the expression of genes of interest. Resulting phenotypes are then used to infer gene function. Unfortunately, RNAi is not without some complications. Empirically, RNAi was shown to function in many different organisms. However, some organisms (*Saccharomyces cerevisiae*, *Trypanosoma cruzei* and *Leishmania major*) are considered to be RNAi-negative, based on the lack of experimental observations for specific knockdown of targeted genes and on the absence of components, such as Dicer and RISC, in the genes of these organisms that are critical for effective RNAi (16,17). More importantly, concern has arisen that the specificity of RNAi, targeted by the sequence of siRNA, may not be perfect. Initially, RNAi was regarded as a highly specific means of gene repression. Several studies dealing with various model systems supported this idea (13,18–20). However, still siRNA can direct RNAi to target mRNA sequences that lack complete sequence identity (21). Agrawal *et al*. (4) forwarded concerns over specificity of gene repression in RNAi. Saxena *et al*. (22) have demonstrated the effect of siRNA mismatches on target specificity in mammalian tissue culture cells and reported 'off-target' gene knockdown. Sequence identity of as few as 11 contiguous nucleotides to siRNA caused direct silencing of non-target genes in experiments conducted on specificity of siRNA in cultured human cells (23). Scacheri *et al*. (24) pointed out that mismatches between siRNA and target sequences could have caused off-target RNAi in mammalian cells but such effects are difficult to detect. Combined, the above examinations of RNAi off-target effects have yielded mixed results. Perhaps as a consequence, RNAi studies do not explicitly control for off-target effects on a routine basis.

Of course, a lack of specificity resulting in knockdown of unknown or unintended genes has considerable negative implications for functional genomics. Target specificity is also of paramount importance when considering applications of RNAi in therapeutics (3,4). For clarification of these uncertainties regarding RNAi, the off-target effect should be evaluated for each gene expressed by the organism under study, by considering multiple possible factors affecting off-target silencing. Such comprehensive studies are most probably expensive and cumbersome to conduct experimentally. A computational approach is less expensive to implement and permits the extension of real parameters into wider ranges for fully observing the trends and effects upon RNAi specificity. This work represents a systematic computational study of RNAi-related off-target effects in several organisms.

Current guidelines for the design of siRNA and dsRNA for RNAi experiments recommend BLAST similarity searches (25) against sequence databases to identify potential off-target genes to improve the likelihood that only the intended single gene is targeted (26). However, the BLAST algorithm was not specifically designed to assess RNAi off-target effects. Therefore, dedicated computational methods were developed for improved detection of sequence identity to accurately and systematically evaluate RNAi off-target effects between siRNA sequences and target genes on a transcriptome-wide scale. In this computational study, three organisms, *Schizosaccharomyces pombe* (fission yeast), *Caenorhabditis elegans* and *Homo sapiens* (human) were examined. The likelihood of off-target effects for all known genes in each of these organisms were evaluated, including factors that may impact the target specificity and efficiency of RNAi. These factors included the length of siRNA, the length of dsRNA, the length of siRNA-target sequence mismatch, the position of mismatch within the siRNA sequence, the position of dsRNA within its target, coding sequences (CDSs) and untranslated regions (UTRs) as targets for RNAi, the chromosomal location and density of genes, and the effect of siRNA selection by rational siRNA design (27). These analyses were aimed to gain insights toward improving specificity of RNAi for functional genomics and potential future therapeutic application by facilitating a better understanding of off-target effects of RNAi. It would also be desirable to include effects such as RNAi directed against promoter regions, concentration dependences and the non-linear silencing effects of siRNA pools. Unfortunately, published empirical data on such effects are currently sparse that we cannot construct a reasonable computational model for them, hence these classes of interactions are omitted from this study.

MATERIALS AND METHODS

Sequence data

The sequence data used in this study were collected from the *S.pombe*, *C.elegans* and *H.sapiens*. RNAi has been observed in each of these organisms and extensive sequence data, including full genome sequences, were available for analysis. These three organisms represent a wide phylogenetic range. We used the cDNA sequences of 5401 genes of *S.pombe* available at the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/yeast/pombe). The cDNA sequences from 22 168 genes of *C.elegans* (release WS110) were obtained from the Wormbase at Sanger Institute. The collective sequence data considered to represent 3′-UTR sequences from *C.elegans* consisted of 1000 UTRs that were present in the expressed sequence tag database combined with sequences that resulted from the UTR prediction method of Hajarnavis *et al*. (28). The dataset of human genes representing 27 852 mRNAs with 3′-UTRs was taken from the RefSeq database at NCBI (http://www.ncbi.nlm.nih.gov).

Modeling RNAi and off-target effects

Although computational methods exist to model aspects of mechanisms that employ short RNA sequences to regulate gene expression, such as microRNA (miRNA) genes (29,30), miRNA targets (31,32) and siRNA efficacy (33–35), none was available to study RNAi off-target effects. Thus,

dedicated computational methods were developed for improved detection of sequence identity to accurately and systematically evaluate RNAi off-target effects based on sequence identity between siRNA sequences and target genes on a transcriptome-wide scale.

RNAi is guided by complete and near complete sequence identity of siRNA and the target mRNA transcript (21–23). siRNA sequences are generated by the activity of Dicer, an enzyme that cleaves long dsRNA into fragments of ∼21 bp (19). To model RNAi, we determined the incidence of sequence identity (exact and allowing for some mismatch) of each of all possible siRNA sequences (arbitrary length range of 17–29 nt) derived from the length of dsRNA (100, 200, 300 and 400 nt starting at the first coding nucleotide, and the sequence region from 100 to 200 nt) representing any of the CDSs relative to all possible siRNA sequences predicted from the CDSs of each of the organisms studied. Sequence identity of the siRNA derived from a given gene by using another gene was considered to signify a potential off-target RNAi. To mimic RNAi that is directed through siRNA with sequence identity to the UTRs of mRNA transcripts, both upstream and downstream UTR sequences (if available) were included for the analysis of off-target effects. With these variables, the effect of length of both siRNA and initial dsRNA upon the chance of off-target effects was investigated. This was implemented as follows.

The similarity between two oligonucleotides is computed with inner product in the feature space using the n-gram feature map, as described previously (36). The use of an inverted file and red black tree (RBT) for calculating the inner products in the feature space achieved efficient computational performance.

Computational representation of siRNA-target binding

We describe each gene by its possible contiguous subsequences of length n (typically ∼21, Table 1 and Figure 1 explain the parameters used in our computational model), called n-grams or n-mers. We consider each gene to be a document described by its coordinates that are indexed into the n-mer space. More formally, each gene g_x in the input space $\mathcal{X}$, consisting of a sequence of characters drawn from the alphabet

Table 1. List of algorithm parameters

Parameter	Description
n	siRNA length (nt)
l	dsRNA length (nt)
pos	Position of dsRNA within target (nt), starting from 5′ end of CDS
m	Length of mismatch permitted (nt)
$mpos$	Position of mismatch within siRNA
u_3	3′-UTR, true when 3′-UTR is included, false otherwise
u_5	5′-UTR, true when 5′-UTR is included, false otherwise
r	Application of rational design rules, true when applied, false otherwise

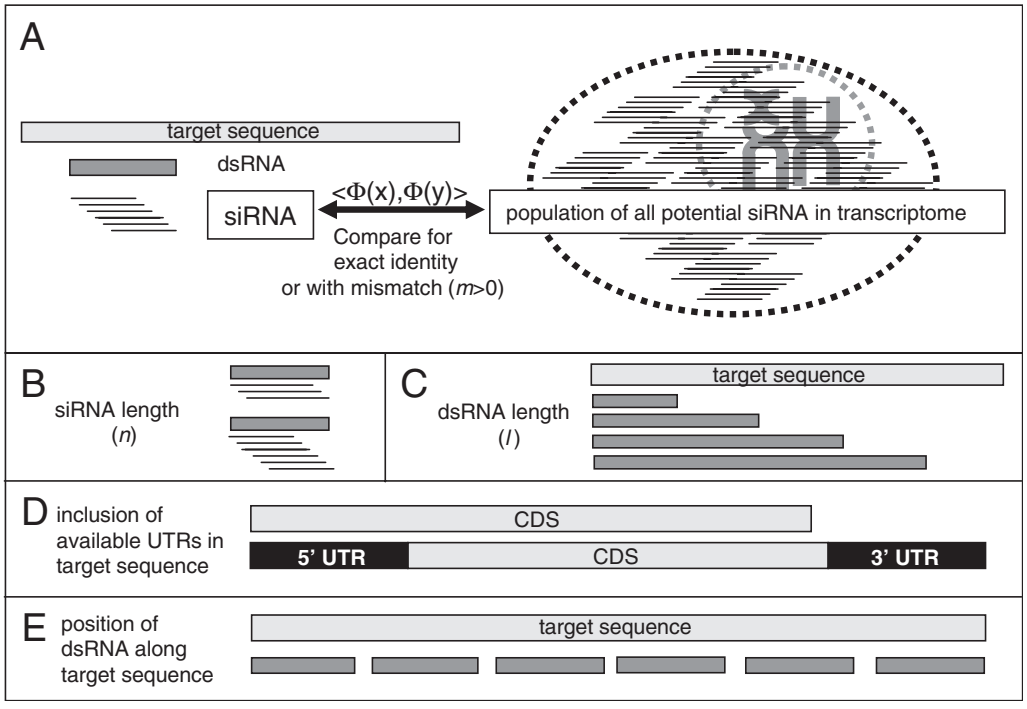

Figure 1. Graphic depiction of variables tested computationally to investigate chance of off-target effect in each of the three organisms (*H.sapiens*, *C.elegans* and *S.pombe*). (**A**) General considerations: a target sequence (representing one particular expressed mRNA) is used as the source of dsRNA of which a pool of all possible siRNA is derived (mimicking the action of Dicer). Each sequence within the siRNA pool was compared for sequence identity (exact: $m = 0$; with mismatch: $m > 0$) to all possible siRNA sequences in the transcriptome through the feature map $\Phi(\cdot)$ to determine chance of off-target errors. The parameters tested are as follows: (**B**) length of siRNA (n); (**C**) length of dsRNA (l); (**D**) addition of available UTR data in the target sequences (u_3 and u_5); and (**E**) position of the dsRNA along the target sequence (pos).

$\mathcal{A} = \{a, c, g, t\}$, $|\mathcal{A}| = 4$, is mapped onto an n-gram feature space, $\mathbb{N}^{4^n}$, by the feature map of exact match

$$\Phi_n^{ex}(g_x) = [\phi_a(g_x)]_{a \in \mathcal{A}^n}, \quad \mathbf{1}$$

where $\phi_a(g_x)$ is the number of times n-gram a occurs in g_x. Therefore, the image of a gene g_x consists of its coordinates in the feature space indexed by the number of occurrences of each of its constituent n-mers. A gene g_y is said to match gene g_x if the following condition is satisfied

$$K(g_x, g_y) = \langle \Phi_n^{ex}(g_x), \Phi_n^{ex}(g_y) \rangle \geqslant T, \quad \mathbf{2}$$

for a predefined threshold T. Then, the similarity measure defined using the inner product in the feature space, $K(g_x, g_y) = \langle \Phi_n^{ex}(g_x), \Phi_n^{ex}(g_y) \rangle$ in Equation 2, actually defines a kernel function that can be used in a support vector machine classifier (37). Here, we use the kernel to match two sequences instead of classification. For modeling RNAi, we choose $T = 1$, since any match between an siRNA and its target mRNA will cause the target to be knocked down. There is evidence that such similarity measures are appropriate models of RNAi off-target effects (23). However, using the feature map in Equation 1 with $T = 1$ establishes a lower bound on cross-reactivity. When more complicated RNAi-binding functions, such as mismatch, wobble and bulge, are modeled, different feature maps could be used than in Equation 1. Changing the value of T can make the similarity measure stronger or weaker. A simple example helps to explain how Equation 2 works. To compute the similarity measure on short sequences O_1 = aacgac and O_2 = aacgtgg using 3mer ($n = 3$) exact match, they are mapped onto the feature space as $\Phi_n^{ex}(O_1) = \{\text{aac}, \text{acg}, \text{cga}, \text{gac}\}$ and $\Phi_n^{ex}(O_2) = \{\text{aac}, \text{acg}, \text{cgt}, \text{gtg}, \text{tgg}\}$. Since 3mers aac and acg occur in both of them, $\langle \Phi_n^{ex}(O_1), \Phi_n^{ex}(O_2) \rangle = 1 + 1 = 2$. Therefore, these two sequences match each other given the parameter and the criterion. For additional details on the kernel function and the computation of the similarity measure, we refer the interested reader to Ref. (38).

Computing the similarity of Equation 2 to find the off-target error in the genome using vector space model directly requires $O(DF4^n)$ time, where F (40×10^6 for *C.elegans* and 60×10^6 for human) is the number of n-grams in the genome that may include UTR sequence and D (close to F) is the amount of n-grams to be compared in the CDSs. For genome-wide scanning, this computing time is prohibitive and can be improved by using the sparsity of the feature vectors. We use an inverted file where the n-grams serve as identifiers and their gene names and positions within the genes serve as attributes (the positions are used for mismatches later). If we ignore n-mers having zero occurrence and allow for the duplication of n-mers, a gene g_x can be represented in the feature space compactly

$$\Phi_n^{ex}(g_x) = \{(a_1, p_1), (a_2, p_2), \ldots, (a_{k_x}, p_{k_x})\}, \quad \mathbf{3}$$

where a_j, $1 \leqslant j \leqslant k_x$, is the j-th n-gram, p_j is its position on g_x and k_x is the number of n-mers in gene g_x with UTRs being optionally included. If the length of g_x is L_x, then $k_x = L_x - n + 1$. In the inverted file, the records for gene g_x contains the triples $\langle a_1, g_x, p_1 \rangle, \langle a_2, g_x, p_2 \rangle, \langle a_3, g_x, p_3 \rangle, \ldots, \langle a_{k_x}, g_x, p_{k_x} \rangle$. The inverted file for the genome of an organism is the collection of the triples of its genes. To speed up computation, we sort the inverted file on the n-mer fields using a RBT, which is a balanced binary search tree and guarantees logarithmic search performance.

$K(g_x, g_y)$ in Equation 2 is computed by searching each n-mer of g_x for g_y in the inverted file. $K(g_x, g_y)$ is the number of occurrence of g_y among the matched genes. Each search in the RBT takes $O(\log F)$ time, resulting in a time of $O(k_x \log F)$ for computing $K(g_x, g_y)$.

Definition of off-target error rate

We define the off-target error using the exact match feature map. However, it is the same for the mismatch feature map defined later. To simulate Dicer's cleavage of dsRNA into siRNAs, we take an oligonucleotide, o_x, as dsRNA from gene g_x and map it onto the feature space, expressed compactly as

$$\Phi_n^{ex}(o_x) = \{(s_1, p_1), (s_2, p_2), \ldots, (s_{l_x}, p_{l_x})\}, \quad \mathbf{4}$$

where s_j, $1 \leqslant j \leqslant l_x$, is the j-th n-mer in o_x and l_x is the maximum number of n-grams in o_x. To obtain the matched genes based on Equation 2, we compute the inner product $\langle \Phi_n^{ex}(o_x), \Phi_n^{ex}(g_y) \rangle$ for each gene g_y in the genome, $1 \leqslant y \leqslant G$, where G is the total number of genes in the genome.

To assess the off-target error rates, we employ measures from information retrieval theory. Let $C_x = \{g_{x_1}, g_{x_2}, \ldots\}$ be the set of genes matched by o_x, not including g_x itself. The precision of a search is the proportion of correct documents to the total number of documents returned by the search. Here, only g_x is correct and the total number of genes returned is $1 + |C_x|$, where $|C_x|$ is the number of genes in C_x. Therefore, the precision of o_x, thus of gene g_x, is $P_x = 1/(1 + |C_x|)$. We define the off-target error as $E_x = 1 - P_x$, so that

$$E_x = \frac{|C_x|}{1 + |C_x|}. \quad \mathbf{5}$$

If C_x is empty, indicating that gene g_x is similar only to itself, then the precision is 100% and the off-target error is zero. The more genes in the set C_x, the larger off-target error silencing gene g_x will have.

We take an oligonucleotide as dsRNA from each gene in the genome and compute its off-target error and average the errors for all genes to evaluate the effects of the parameters. Thus, we define the average error rate for a given parameter set to be

$$E(\Theta) = \frac{1}{G} \sum_{i=1}^{G} E_i, \quad \mathbf{6}$$

where E_i is the error for gene g_i and Θ is the set of parameters

$$\Theta = \langle l, pos, n, m, mpos, u_3, u_5, r \rangle. \quad \mathbf{7}$$

In the above expression, n is siRNA length in nucleotides (nt). The length of siRNA produced by Dicer appears to vary slightly across organisms (4), so we examine the effect of siRNA length on off-target error rate. A computational approach is also able to simulate siRNA lengths that do not occur *in vivo*, allowing us to assess the tradeoffs between siRNA lengths and off-target error rates. l is dsRNA length (nt). As dsRNA can be synthesized and introduced through a short hairpin experimentally as an alternative to using siRNA directly, its length determines the number of possible siRNAs

produced by Dicer. By varying l, we can investigate the chances of off-target error using different dsRNA lengths. *pos* denotes the position of a dsRNA on its target mRNA, measured in nucleotides from the 5′ end of the CDS. Different segments of sequences on a target gene have varying silencing efficacies and Dicer has been reported to preferentially cleave siRNA on its target (26,39). To study whether dsRNA position has an effect on its potential of off-targeting, we examine dsRNAs out of different positions from their target gene. m is the mismatch length in nucleotides. Experiments have shown that RNAi works despite the existence of mismatches between the siRNA and its target (22,23). To observe the effect of mismatch on off-target error rate, we use a range of mismatches in our computational procedures. *mpos* is the position of the mismatch within the siRNA. Differential silencing efficiency among variable mismatch positions has been reported in biological experiments, demonstrating mismatches in certain regions within the siRNA is critical for effective knockdown (22). We examine the positional effect of mismatch by changing *mpos*. In order to control the position of the mismatch, we only consider contiguous mismatches. They are also most frequently experimented. u_3 and u_5 indicate the inclusion of the 3′-UTR and 5′-UTR, respectively. It is possible for an siRNA to bind their targets by matching to their UTRs (4). We study the effect of UTR on the off-target error by optionally including them. r indicates the application of siRNA selection by rational design. siRNA selection using rational design rules (27) can effectively increase silencing efficacy and reduce the number of siRNAs used. We investigate the effect of rational design on off-target by optionally applying these rules. Figure 1 and Table 1 illustrate these parameters. The average off-target error represents the chance that siRNAs from a dsRNA of a gene will cross-hit some other genes. When we know the average error $E(\Theta)$ for a particular combination of parameter Θ, we can find $Z(\Theta)$, the average number of genes that will be targeted incorrectly by targeting each gene

$$Z(\Theta) = \frac{E(\Theta)}{1-E(\Theta)}. \quad \mathbf{8}$$

The parameters in Equation 7 are based on the RNAi experiments but extended to include much wider ranges. This extension will facilitate our quantitative study. The calculations of Equation 6 are conducted through searches in the inverted file and can be performed in $O(D\log F)$ time.

An algorithm for detecting siRNA-target binding allowing mismatches

Experiments have shown that RNAi works despite the existence of a number of mismatched nucleotides between the siRNA and its target gene (22,23). However, the efficacy changes with the length of the mismatch and the position of the mismatch on the siRNA. Several algorithms have been developed for string mismatching, a problem that relates to siRNA-target similarity. Leslie *et al*. (36) used a trie to construct a mismatch tree for computing their mismatch string kernels applied in a support vector machine classifier to detect protein families. Suffix trees were used as data structures to predict putative RNAi (40). Amir *et al*. (41) have developed an algorithm for single mismatch string searches ($m = 1$), which is not enough for our study. BLAST (25) also allows for mismatch by using substitutions based on alignment cost. However, the related mismatch algorithms are not particularly developed for RNAi and cannot control the positions of the mismatch as required in computational models for RNAi. We define mismatch feature map as follows.

For an n-mer a from an alphabet $\mathcal{A}$, define its mismatch neighborhood $N^{\text{mis}}_{m,p}(a)$ as all n-mer γ from $\mathcal{A}$ that differ from a by at most m mismatches starting at position p in a and ending at position $p + m - 1$ in a and $\Phi^{\text{mis}}_{m,p}(a) = [\phi_\gamma(a)]_{\gamma\in\mathcal{A}^n}$, where $\phi_\gamma(a) = 1$ if $\gamma \in N^{\text{mis}}_{m,p}(a)$ and $\phi_\gamma(a) = 0$, otherwise. The feature map of a gene g_x is defined as the sum of the feature maps of its n-mers

$$\Phi^{\text{mis}}_{m,p}(g_x) = \sum_{a\in g_x} \Phi^{\text{mis}}_{m,p}(a). \quad \mathbf{9}$$

The mismatch kernel is computed by replacing the exact feature map in Equation 2 by the above mismatch feature map. To compute the average error in Equation 6 with a mismatch of m nucleotides long, a straightforward search would require 4^m exact match searches, and takes a time of $T_{sf} = O(4^m D\log F)$. If each exact match search needs 30 min, then a 3-nt mismatch search would need 32 h. To improve computational performance, we developed a novel flexible and efficient algorithm for the mismatch searches in RNAi that allows for arbitrary length of the mismatch and arbitrary positions of the mismatch in the middle of an siRNA. Our algorithm manipulates reverse sequences [as by Amir *et al*. (41)] but uses different data structures and search strategies. This algorithm is more computationally efficient than the mismatch tree algorithm proposed by Leslie *et al*. (36) with respect to time and space usage.

We first introduce some notations. Let $S = \{s_1, s_2, \ldots, s_N\}$ be a set of strings of length k from an alphabet $\mathcal{A}$. Suppose string $s_i = a_1 a_2 \cdots a_k$, where $a_j \in \mathcal{A}$, $1 \leqslant j \leqslant k$, has reverse string $\bar{s_i} = a_k a_{k-1} \cdots a_1$. We next define mirrored tree and leading range used in our algorithms.

DEFINITION. *A mirrored tree of a binary search tree* (*BST*) *populated with strings from S is the BST populated with reverse strings* $\overline{s_1}, \overline{s_2}, \ldots, \overline{s_N}$. *A u leading range of a string s from S searched in a BST is the set of nodes returned by a search that only matches the beginning u letters of s*.

The mismatch kernel corresponding to Equation 9 can be computed by the mirrored tree search (MTS) in Algorithm 1. We omit its correctness proof owing to space limitation.

Algorithm 1. Mirrored Tree Search, MTS (n,m,p)

```
Build RBT T_1 and mirrored RBT T_2 for the inverted file using n-mer
for each gene g_i in the genome do
  Take a subsequence d_i in g_i
  for each n-mer s_j^i in d_i do
    Get R_1, the p − 1 leading range of s_j^i from T_1
    Get R_2, the n − m − p + 1 leading range of the reverse of s_j^i from T_2
    Find C_j^i = R_1 ∩ R_2
  end for
  Calculate off-target error E_i for g_i using C_i = ∪_j C_j^i
end for
Calculate average off-target error for the genome
```

At Steps 5 and 6, the substring before the mismatch is exact-matched in T_1 and the leading range is stored in R_1, the substring after the mismatch is exact-matched in T_2 and the leading range is stored in R_2. The genes corresponding to the mismatch letters are sandwiched in C at Step 7 by the intersection based on gene names and positions of the n-mers. At Step 9, E_i is computed using the definition based on C_i.

Let the size of the inverted file be F and the total number of n-grams from all the dsRNAs be D. MTS has a cost of $T_{mt} = O(D(2\log F + C))$, where $O(2\log F)$ is the search time in the two RBTs and C is the cost of obtaining the leading ranges and the intersection. Using proper join algorithms C can be bounded by $O(|R_1| + |R_2|)$ (42). Empirically, C is small and treated as a constant. The mismatch tree algorithm (36) has a complexity of $O(DLn^m 4^m)$, where L is the average length of genes. Since usually $F < 10^8$ and $\log F < 30$, MTS is much faster. MTS's speed-up over the straightforward method is roughly $\mathrm{Sp} \approx O(4^m)$. Empirical results demonstrated that MTS achieved speed-ups of two-orders of magnitude on average for the three organisms. However, it uses more space because of the mirrored tree.

Simulating positional effect of mismatches

Experiments suggested that nucleotides in the region of 2–9 nt at the 5′ end of the guide strand are crucial for gene silencing (22,43). It therefore seems that transcripts containing sequence identity within this critical binding region would have a higher probability of being targeted for silencing and that mismatches within this region would have a more significant effect on reducing off-target silencing. To see this positional effect of mismatch, we use a weighted scheme where we assign lower silencing efficiency scores if the mismatches are in the critical binding region, and higher scores if the mismatches are outside of the region. The silencing efficiency score for silencing a gene is the sum of all the scores contributed by each siRNA. A gene is considered silenced only when its total efficiency score is above a threshold. We use contiguous mismatches to control their positions.

Distribution of redundant siRNA sequences in the transcriptome

A coincidental high frequency of particular 21mer target sequences within a transcriptome would increase the probability for off-target effects of any given siRNA sequences. The transcriptome of each organism tested was described as a collection of all possible 21mer sequences contained within and frequency of each sequence was determined. Web utilities were made available so that the frequencies of siRNAs of a particular sequence and genes targeted by the sequence can be retrieved, for the benefit of siRNA design.

Effect of dsRNA position

Frequently, dsRNA (of various lengths) is used to affect RNAi experimentally. Success was obtained with dsRNA sequences from various locations within the full-length CDS of targeted genes. The position of dsRNA along the target sequence was investigated as a parameter for off-target effects. Beginning with the first nucleotide of the coding region of a gene, the start position of dsRNA was incremented 6 nt (two codons) until position 600 (on average, the final dsRNA closely approached the end of the CDS). The off-target error (based on exact sequence identity) was determined for such dsRNA for all CDSs in the transcriptome.

Off-target error distributions in chromosomes

The physical distribution of genes on chromosomes is not uniform. Often more genes are located in the middle of a chromosome than in the ends. The chance for off-target errors, based on exact sequence identity, for each gene was plotted against its coordinates on the genetic map of *C.elegans*. The curves were smoothed by averaging the error of a gene with that of its neighbors.

Implementation of rational siRNA design

Different siRNAs from the same target gene have highly variable efficacy (27,43). Several biological and thermodynamical properties have been identified to characterize siRNA sequences that mediate especially efficient RNAi knockdown (27,34,35,43,44). This has led to a set of rules for 'rational design' to optimize siRNA development (27). All possible siRNA sequences of each of the organisms studied were scored by the eight criteria of rational design. The length of siRNA sequences analyzed ranged from 17 to 29 nt (odd numbers only). The off-target error (based on exact sequence identity only) was determined for the highest scoring siRNAs from each gene (in pools of 5, 10 and 20 sequences) to evaluate whether rational design may reduce off-target errors.

RESULTS

Off-target error based on siRNA sequence identity versus the transcriptome

The comparison of siRNA sequences with an arbitrary range of lengths (from 17 to 29 nt) derived from particular dsRNA sequences against all possible targets on a genome-wide scale disclosed that the length of siRNA sequences is an important parameter for determining off-target effects as defined by sequence identity with other than the intended target sequence for all three organisms tested. Only CDSs were used as target sequences. The chance for off-target errors decreased with increasing lengths of siRNA. siRNA of 21 nt proved optimal, the chance for off-target effects with this length was significantly lower than for shorter siRNA sequences whereas it did not differ significantly from that of longer siRNAs (Figure 2). The off-target effects increased when longer dsRNA sequences (from 100 up to 400 nt) were used to generate the pool of siRNA. The sequence diversity inherent to dsRNA increases with length, leading to a more diverse pool of siRNA sequences. However, the optimal length for siRNA remained at 21 nt. When siRNAs are short, longer dsRNAs generate larger off-target errors. When siRNAs are relatively long, longer dsRNAs cause slight increase in the error. However, the effect of dsRNA length is more drastic in larger genomes (such as *H.sapiens*, Figure 2A) than in smaller genomes (as in *S.pombe*, Figure 2C). Overall, the larger genomes had higher chances of off-target error involving other genes than the small genomes (*S.pombe*), as compared in Figure 2D. In the case of RNAi in *H.sapiens*, an off-target

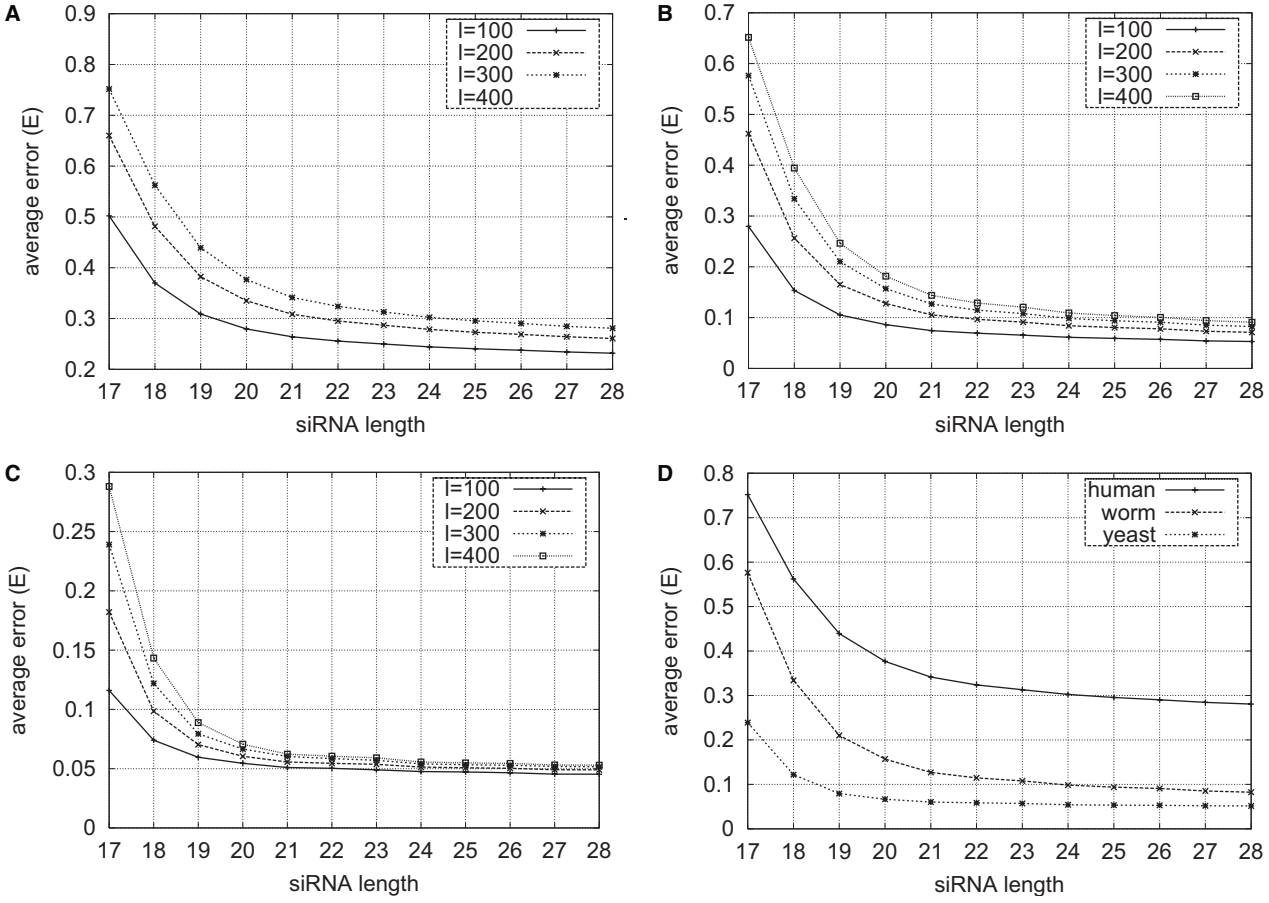

Figure 2. Effect of siRNA length and dsRNA length on off-target error rates under an exact match ($m = 0$) siRNA homology function. (**A**) *H.sapiens* and (**B**) *C.elegans*. (**C**) *S.pombe*, $l = 100$–400, $n = 17$–28. (**D**) Comparison of the three organisms for $l = 300$, $n = 17$–28 and $m = 0$. Within the length range tested, siRNA of 21 nt optimally combined target specificity and minimum length. Chance for off-target error increases significantly for smaller siRNAs, yet is not significantly different from longer siRNAs. Increased length of dsRNA yielded more diverse pools of siRNA, resulting in increased chance for off-target errors.

error of 30% is equivalent to 0.43 non-targeted genes, based on Equation 8.

$$Z = \frac{E}{1-E} = \frac{0.30}{1-0.30} = 0.43. \qquad \mathbf{10}$$

Effects of length and position of mismatches

Allowing sequence mismatch of up to nine contiguous nucleotides between siRNA and its target sequences increased the off-target error (Figure 3). As shown, off-target error increased dramatically with longer mismatches.

Using the weighted scheme for simulating the positional effect of mismatches, we found that off-target error rates corresponding to mismatches within the critical binding region (2–9 nt at the 5′ end of the guide strand) were significantly lower, whereas the error rates corresponding to mismatches outside this region were much higher, consistent with the findings in the literature (22,23,43). Figure 4 shows the positional effect of mismatches in *C.elegans*, *S.pombe* and human genomes, for the cases of $n = 21$, $l = 100$, $m = 3$ and $mpos = 1$–19. As shown in this figure, the off-target errors with mismatches in the critical binding region are significantly lower, with *P*-values close to zero for the three organisms. In the case of *C.elegans*, for example, the average off-target error with mismatches in the critical binding region was 10.2%, whereas the average error with mismatches outside the region was 15.8%, with a standard deviation of 0.41% and *P*-value ≈ 0.

Effects of UTRs

Incorporation of available (not for *S.pombe*) UTR sequence data considerably increased the size and diversity of the target sequences for *H.sapiens* and *C.elegans*. The 3′-UTR sequences described for human transcripts when added to the inverted file containing the CDSs, increased the RBT by 58%, and the number of leave nodes grew from 41.4 million to 65.5 million. The use of exact sequence identity as parameter while analyzing siRNA of various lengths, derived from different lengths of dsRNA representing CDS target sequences only, showed only non-significant increase in off-target errors of RNAi in the case of *H.sapiens* and *C.elegans* (Figure 5). For example in *H.sapiens*, the average error over siRNA length from 17 to 28 is $\sum_{n=17}^{28} E(l = 200, u3 = 0, n)/12 = 0.340$, when $l = 200$ and 3′-UTR sequences were not considered; whereas inclusion of 3′-UTR yielded an average error of $\sum_{n=17}^{28} E(l = 200, u3 = 1, n)/12 = 0.353$. These average errors are not significantly different (*P*-value = 0.81).

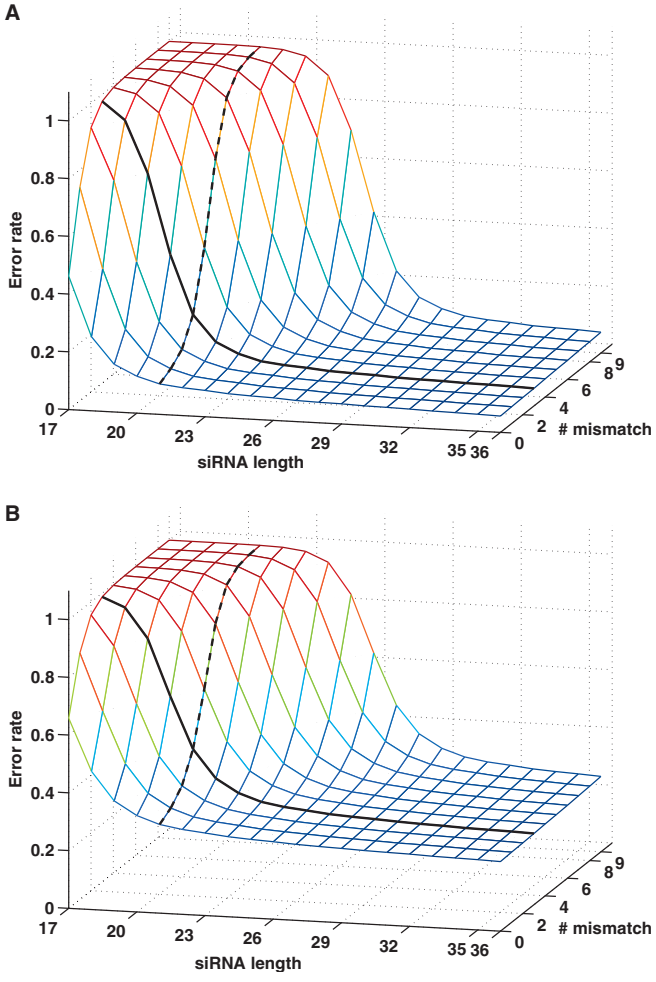

Figure 3. Effect of mismatch on RNAi off-target error rate in (**A**) *C.elegans* and (**B**) *H.sapiens*. Allowing for mismatches significantly increased the off-target error for siRNA of various lengths derived from CDSs of the transcriptome. Results shown were derived for $l = 200$, $n = 17$–35 and $m = 0$–9. Dashed curves indicate the positions of the planes of $n = 21$ and solid curves indicate the positions of the planes of $m = 3$ in the three-dimensional plots.

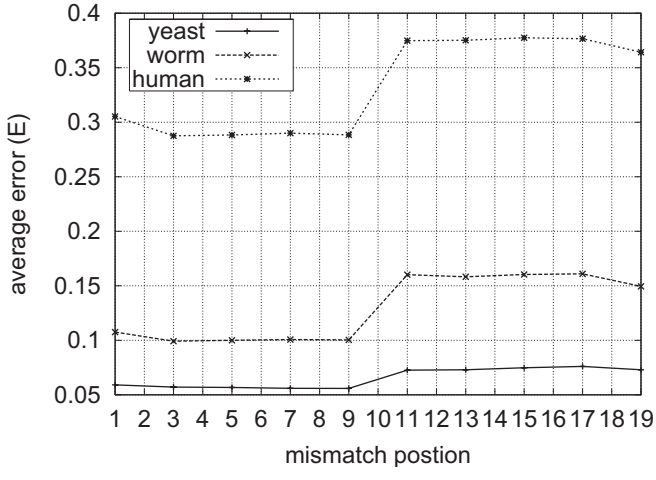

Figure 4. Effect of mismatch positions in *H.sapiens*, *C.elegans* and *S.pombe*. Mismatches in the region of 2–9 nt at the 5′ end of the guiding strand reduced off-target chances. Results shown were derived for $n = 21$, $l = 100$, $m = 3$ and $mpos = 1$–19.

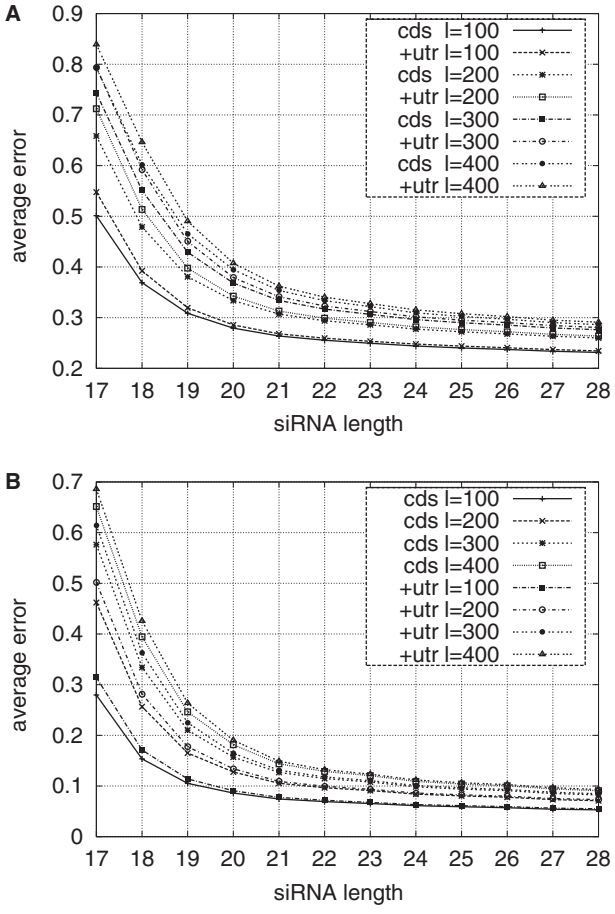

Figure 5. Effect of including UTR sequences with the CDS data for RNAi off-target error in (**A**) *H.sapiens* and (**B**) *C.elegans*. The off-target error for siRNA (various lengths) derived from CDSs of the transcriptome did not change significantly when available UTR sequence data were included among the sequences considered as the transcriptome (average *P*-value = 82%); $l = 100$–400, $n = 17$–28 and $m = 0$. CDS indicates the case where UTR is not included; '+UTR' indicates the case where 3′-UTR is included.

Frequency of specific 21mer sequences in different transcriptomes

The sequence data of the transcriptome of each of the three organisms studied were computationally scanned for the occurrence of all possible 21mer sequences representing siRNA, derived from the same transcriptome. Particular sequences were present at distinctly different frequencies (Figure 6). This indicates that a direct sequence comparison analysis of potential siRNA sequences versus the transcriptome of specific organisms can identify siRNA designs with reduced chance for off-target error owing to a low frequency of potential off-target sequences in the transcriptome. Table 2 displays the proportion of unique siRNAs in the organisms. It indicates that out of all possible siRNAs in human 83.1% are unique, in *C.elegans* 86.6% are unique, and in *S.pombe* 99% are unique.

To assist siRNA designers to evaluate off-target errors, we have made available the frequency count of each siRNA in the three genomes of *H.sapiens*, *C.elegans* and *S.pombe* on the Web at http://rnai.cs.unm.edu/rnai/off-target/sirna_freq/. The website accepts siRNAs and returns their occurrence count

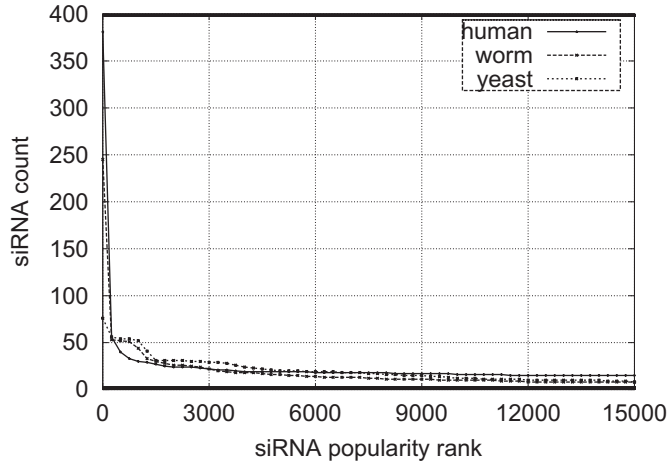

Figure 6. siRNA frequency distribution. The *x*-axis is the popularity rank of the siRNAs (the most frequently occurred siRNA is the most popular one and has the smallest abscissa). The *y*-axis is the corresponding count with which the siRNA occurs in the genome. The curves have very long and flat tails and the *x*-axis is actually only 1000th of the total data.

Table 2. siRNA uniqueness

Organism	ϕ (%)	Max count	Average count	ρ
Human	83.1	381	1.29	0.0024
Worm	86.6	245	1.19	0.017
Yeast	99.1	141	1.02	0.0016

ϕ is the percentage of unique siRNAs. The third and the fourth columns show the maximum and average counts with which the siRNA occurs. ρ is the correlation coefficient between the frequency count and the rational score of an siRNA.

that serves as indicators for off-target chances. We also provide a web tool that searches for the genes targeted by a given sequence allowing mismatches and different siRNA lengths (http://rnai.cs.unm.edu/rnai/off-target/genes-targeted/).

Effect of dsRNA position along the target sequence

The incremental variation of the position of the dsRNA (that served as source for the siRNA) along the target sequence showed that off-target errors were significantly lower for the beginning 100 nt positions than for those in the following positions. Figure 7A shows this position effect for *C.elegans* genome for the cases of $l = 100$, 200 and $n = 22$. The off-target error in these two segments were significantly different, *P*-value $< 10^{-3}$. The same phenomenon is evident for *H.sapiens* (Figure 7B) and *S.pombe* (Figure 7C). However, the differences are smaller and become insignificant for most of the parameters for *H.sapiens*. The initial region of human-derived CDSs associated with lower off-target error is decreased in length when the length of the dsRNA was increased.

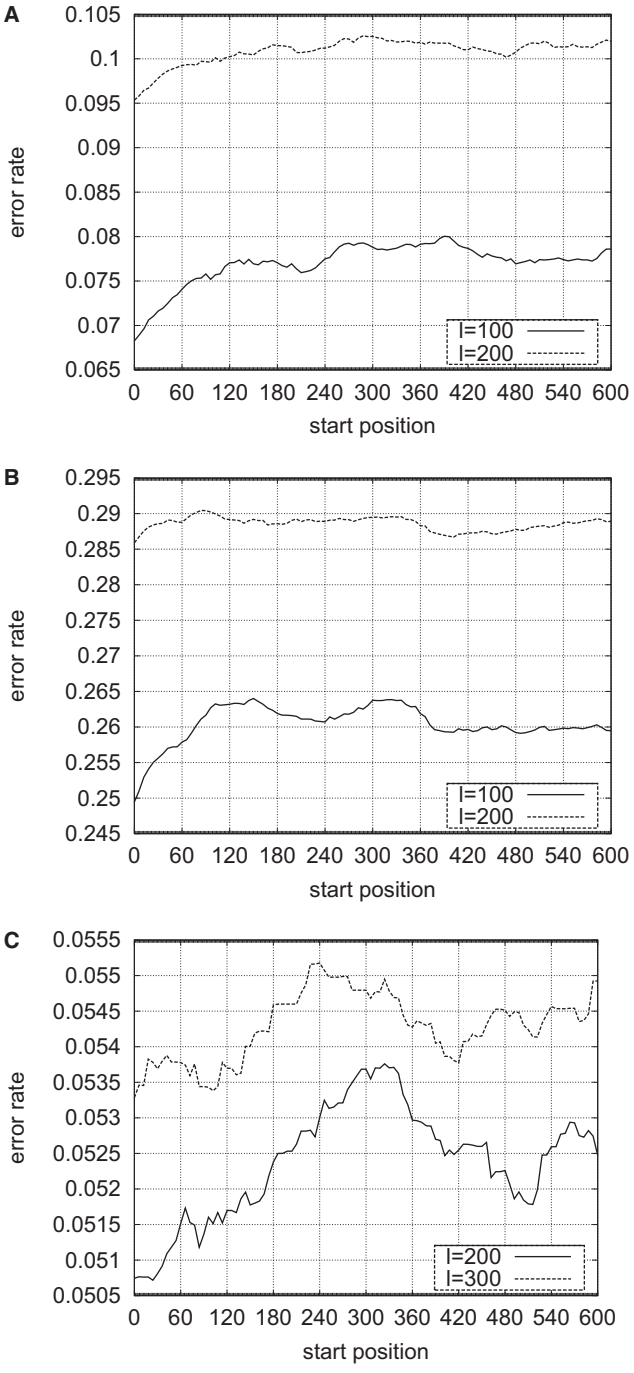

Figure 7. RNAi off-target error relative to varied position of the dsRNA along the target sequences. The off-target error is significantly lower when targeting the beginning 100–140 nt by dsRNA for generating siRNA. The *x*-axis is the start position of the dsRNA, *pos*. The *y*-axis is the error rate. (**A**) For *C.elegans*, off-target error shown for dsRNA lengths $l = 100$ and 200 nt, length of siRNA $n = 22$ and position *pos* 0–600. (**B**) Off-target error as a function of dsRNA position effect in *H.sapiens* ($l = 100, 200, n = 25$). (**C**) For *S.pombe* $l = 200, 300$, $n = 25$. When $pos \approx 100$, the error reaches a local minimum.

Off-target error distributions on physical maps of chromosomes

The chance for off-target errors for each CDS in the transcriptome was mapped onto the physical map of chromosomes from *C.elegans* (Figure 8). The physical maps show that genes are distributed in a rough bell shape around the density center of each chromosome. The distribution of off-target errors for RNAi differed considerably; the off-target errors are low in the gene-dense centers of chromosomes I, II, III and X. In all the chromosomes except for chromosome I, there

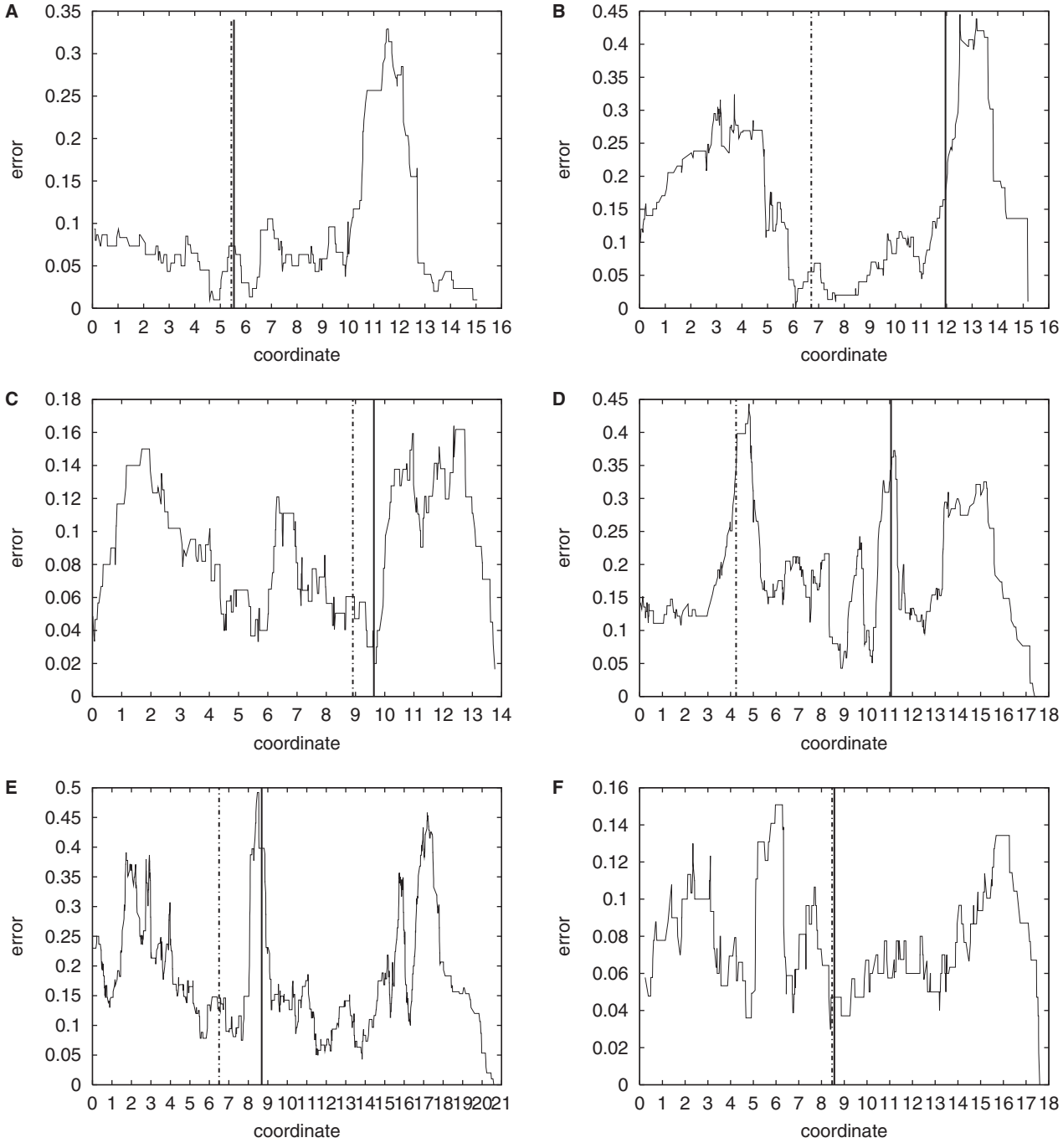

Figure 8. The distribution of RNAi off-target error of each gene plotted against the position of genes on the physical map of chromosomes of *C.elegans* (A–F represent chromosomes I–V and X, respectively). These plots disclosed that the chromosomal position is predictive of the off-target error rate. The curves were smoothed by averaging the off-target errors between neighboring genes. Dashed bars indicate physical center of the chromosomes (cM = 0), solid bars identify the position where most genes are centered (density center). Parameters are $l = 300$, $n = 21$, $mpos = 0$. Unit on the x-axis is million nt.

are two high-error regions in the low and high telomeric ends of chromosomes apart from the centers. In chromosome IV, the error is high in the density center region, but lower than the two peaks off the center. In chromosome V, the high off-target error rate in the central gene-dense region also exceeds the error rate of two regions that are off the center. The genes in the coding-dense region of a chromosome did not necessarily have the highest off-target errors. These error distributions suggest that genes in the high density regions may use more unique sequences than in regions sparsely populated by genes. In chromosomes II, IV and V, the peak off-target errors were close to 0.5. Thus, based on Equation 8, targeting each gene in these regions by RNAi is on average associated with a potential off-target knockdown of one other gene. Thus, RNAi experiments to knockdown genes in such regions may consider more additional or even specific controls.

Effect of rational siRNA design

All possible siRNA sequences of various lengths from dsRNA (l = 300) were selected using rational design parameters to identify a subset of siRNA sequences that are more likely to effectively guide RNAi. The off-target error for this subset of sequences was determined for *H.sapiens* and *C.elegans* owing to their high off-target error rates (Figure 9). This approach reduced off-target errors substantially. The algorithms used in this study were easily modified to incorporate the rational design filter, including only high-scoring siRNAs (score = 6 out of 8) (27) for the off-target error analysis. The average off-target error of all (non-rational design) siRNAs was significantly larger (P-value = 0.001) than those of siRNA sequences selected by rational design. The average off-target error rates determined using rational design pool of sizes of 5, 10 and 20 (representing siRNA sequences that scored highest), and siRNAs selected as scoring 6 or higher by the rational design filter were not significantly different, P-value = 0.13. Thus, rational design reduced the off-target error from 34 to 24% (n = 21, l = 300, *H.sapiens*, Figure 9A).

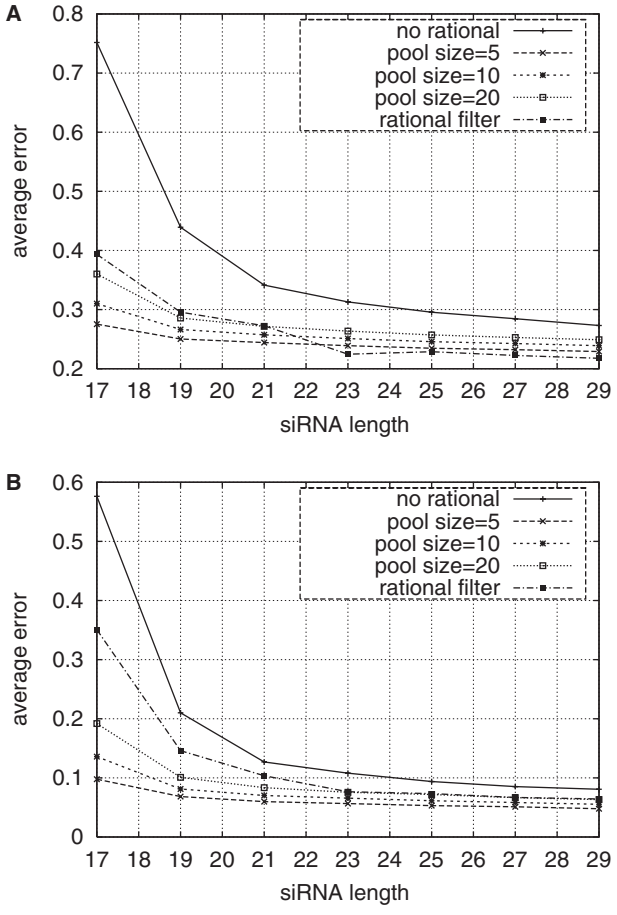

Figure 9. Effect of rational siRNA design in *H.sapiens* and *C.elegans*. (**A**) *H.sapiens* and (**B**) *C.elegans*. Results shown for l = 300, n = 17–29 (only odd number of lengths are considered since rational design rules need to compute the center of an siRNA), pool size = 5, 10, 20 and rational filter is also considered. The selection of siRNA with properties that were empirically found to be associated with highly functional siRNA sequences reduced off-target error relative to that of non-filtered siRNA. Note that off-target error is never reduced to zero.

Significant reduction of off-target errors was also observed for *C.elegans* (Figure 9B).

To understand the relationship of an siRNA's frequency in the genome and its efficacy represented by its score of rational design, we computed the correlation coefficient between the count and the rational score of each siRNA. These correlation coefficients are very small, as shown in Table 2, indicating that the frequency and the rational score are not correlated. This independence between frequency of an siRNA and its rational score suggests that the objectives of minimizing off-target error rates of siRNAs and maximizing their efficacy can be pursued independently in an siRNA design.

DISCUSSION

This computational study of mechanistic aspects of RNAi against the background of extensive transcriptome and genome information available in the nematode *C.elegans*, *H.sapiens* (human) and (to a lesser extent) for *S.pombe* (fission yeast), indicated a considerable likelihood that the specificity of RNAi knockdown is compromised by off-target RNAi effects. The similarity of observations from organisms of a wide phylogenetic range (fungi to both protostome and deuterostome animals) suggests that the conclusions from our analyses may provide insights into general aspects of RNAi. The results reported here were derived from computational approaches only; the feasibility of experimental validation is compromised by the large (genome-size) scale of the sequence data considered in these analyses. However, the parameters used for the computational analyses were applied at a high stringency compared with the conditions that allow RNAi *in vivo*. For instance, only sequence identity and minimal mismatch were considered to define siRNA specificity for a target sequence, bulge or wobble phenomena that relax sequence-specific target recognition by siRNA (22) were not allowed for. In addition, with a computational approach it was feasible to test parameters (such as lengths of dsRNA and especially siRNA) beyond the naturally occurring ranges to examine properties and trends of RNAi specificity. Our work does, however, suggest an empirical investigation that would be informative about both *in vivo* off-target rates and the properties of the siRNA binding/knockdown process. In each of the organisms studied here, we have identified a number of siRNA that have the highest potential for off-target effects, along with the predicted affected genes and predicted efficacy according to the rational design rules (http://rnai.cs.unm.edu/rnai/off-target/). An *in vivo* study of the knockdown produced by some or all of these siRNAs with regard to the putative affected genes and controlled by monitoring predicted non-target genes (measured, e.g. by microarray analysis) could reveal whether the predicted off-target effects do, in fact, occur in living systems. The rates of off-target knockdown would help to calibrate the predicted rates in this paper. Furthermore, such an experiment would contribute evidence toward the current debate of whether or not efficacy is purely a function of the siRNA or is also dependent on the target molecule (45–47).

The algorithm used to detect sequence similarity as parameter for off-target RNAi was designed specifically for use with short (siRNA) sequences, while it also incorporated the

use of dsRNA as a source for siRNA sequences. Thus, the analyses are relevant for two ways to experimentally affect RNAi, introduction of dsRNA and siRNA (18,26). The algorithm is superior to the BLASTN algorithm (25), which is usually recommended to evaluate potential off-target properties of siRNA and dsRNA designs toward other genes (26). While BLAST search offers some protection against off-target effects, and is certainly better than no control whatsoever, it is not, by itself, sufficient for general use for at least two reasons. First, the BLAST homology function was not particularly developed to model the RNAi-binding process and does not account for some of its known features. For example, mismatches and bulges are known to have differential effects on efficacy, varying along the length of the siRNA (22,23). Although BLAST allows for mismatch, insertion and deletion based on alignment cost, it cannot control the positions of these imperfect match patterns. Our algorithm is capable of modeling these patterns by controlling the length and positions, allowing it to detect off-target effects that would be missed by BLAST searches. Second, BLAST is suitable only when the entire genome sequence is available—in the absence of complete genome information, it is possible that significant off-target interactions will be missed. Although we can also only search complete genomes in the current work, we have quantified expected off-target error rates in a number of organisms, establishing a range of probable off-target rates. In an otherwise unsequenced organism, these bounds can be used to estimate the probability of off-target effects based on comparison of its genome size and evolutionary history. They can also be used to ameliorate such effects through multiple trials with varying siRNA selected from the target gene. Using off-target framework built in this work, we are able to develop quantitative models to predict off-target errors by incorporating a number of variables such as genome size and chromosomal location of a target gene in addition to the parameters we used in Materials and Methods. These models will provide reliable prediction of false positive error rates when an organism is partially sequenced.

We should note that our predictions neither include the effects of siRNA concentration nor do they attempt to account for the non-linear (synergistic or mutually interfering) interactions of a pool of siRNA. It is clear that both these effects are of critical practical consequence and that a computational model supporting them is desirable. At the moment, however, there is insufficient published data on the efficacies of pools to be able to construct a high-confidence model of pool effects. From some reports (15) it is clear that simplistic models, such as linear combinations weighted by concentration, are inadequate. Thus, the results in this paper do not attempt to model either concentration or non-linear siRNA pool effects. Our results should, therefore, be interpreted as the chance that any single siRNA arising from a chosen dsRNA has a chance of off-target interaction within the genome. In practice, this may be an overestimate of true off-target effects, but it does still provide an indication of off-target genes that should be monitored for potential off-target repercussions.

Remarkably, the examination of RNAi off-target error as a function of siRNA length disclosed that siRNA sequences of 21 nt, the length most observed *in vivo*, optimally balanced target specificity and low chance of off-target RNAi. siRNA sequences of <21 nt had increased chance for off-target effects whereas longer sequences did not gain adequate target specificity to significantly reduce off-target reactivity. This siRNA length effect suggests that the chance for off-target RNAi effects may increase with the use of artificial siRNA sequences of <21 nt, such as 12–15 nt dsRNA fragments that result from RNase III digestion of dsRNA (48). The protozoan parasite *Trypanosoma bruci* employs comparatively long siRNA (24–26 nt) to target RNAi (39), perhaps for the benefit of gaining some critical specificity of RNAi. However, sufficient sequence data are lacking at this time to validly investigate the off-target dynamics for siRNA of various lengths in this organism.

Despite inherent properties that combine optimally for specific sequence-based recognition, 21 nt siRNA still have a considerable chance for off-target effects when considering all coding domains within a transcriptome. Not surprisingly, the incidence of off-target effects increased when sequence mismatch of up to nine consecutive residues between the siRNA and the potential targets was allowed for. Varying the position of these mismatches within the siRNA sequence changed the number of potential target sequences. Consistent with experimental observations (22,23,43), we found that off-target error rates corresponding to mismatches within the region of 2–9 nt at the 5′ end of the guide strand were significantly lower.

The off-target effects also increased following inclusion of upstream and downstream UTR sequences within the target sequences, to reflect the *in vivo* reality that complete mRNA transcripts (not just the protein-encoding sequences) can be attacked by RNAi. Although this increase was not significant (Figure 5), this result suggests that the nucleotide usage in UTRs substantially differs from that in coding regions. Finally, off-target errors increased when using longer versus shorter lengths of dsRNA to generate the siRNA population. Intriguingly, our methods showed that dsRNA representing the region of the first 100 nt of the 5′ terminus of CDSs yielded the lowest chances for off-target effects. This particular region may differ between genes for proteins that function intracellular versus proteins that are released extracellularly. The 5′ sequence of the latter category of genes encodes for signal peptides or membrane anchors. The specific sequence constraints that ensure functionality of these domains (49), may subdivide the transcriptome into smaller populations of target sequences. Although the beginning of the CDS has lower off-target error, this region is not recommended for dsRNA design because it is rich in regulatory protein-binding sites (26). There is also an empirical evidence that Dicer produces more siRNA toward the 3′ portion of the target gene (39,50). In all, the reduction in off-target error was not significantly different for dsRNA from the first 100 nt versus dsRNA representing residues 100–200 nt of CDSs.

Combined, the above computational findings suggest an extensive potential for off-target effect of RNAi experiments. However, in practice, chances for off-target errors may be less severe. RNAi targets mRNA for destruction and can only knock down genes that are expressed when siRNA is present. Potential off-target genes (that have adequate sequence identity to siRNA) will not be affected if they are not expressed simultaneously with the intended target gene. Our analysis showed that relatively few siRNA targets a sequence that is repeated frequently throughout the transcriptome of each of

the organisms tested. In fact, siRNA designs can be screened for this property (http://rnai.cs.unm.edu/rnai/off-target/) to avoid the use of siRNA with increased chance for off-target errors. Moreover, we determined the chance for off-target error for each gene within the transcriptome of *C.elegans* relative to its position on the physical map of the genome of this nematode. CDSs from chromosome regions that contain more densely packed genes had a lower probability for off-target RNAi, as observed in all chromosomes except chromosomes IV and V. This implied that densely packed genes generally employ more unique sequences within the genome of *C.elegans*. Regardless, once a physical map is available for an organism, it may be possible to correlate the need to consider RNAi off-target error for a particular gene with the location of that gene within the genome. In addition, the results of the combined analysis suggest a trend where the chance for the off-target error is elevated for larger genomes. Of note, *C.elegans* and *H.sapiens* have roughly the same proportions of unique siRNAs (Table 2), but the off-target error rate in *H.sapiens* was much higher (Figure 2D). Sequence comparison showed that transposable elements may not be the source of these frequent 21mers, and the true origin remained to be determined.

Finally, several properties of siRNA sequences have been found to be associated with a high efficacy to cause RNAi. For instance, the relative thermodynamical stability of the sequence termini may determine how a double-stranded siRNA dissociates to correctly incorporate the negative RNA strand into the RISC complex (43). Such properties have been combined into rational design methods for improving the siRNA efficacy (43). Implementation of rational design yielded a considerable reduction in the number of functional siRNA sequences derived from the transcriptomes of *H.sapiens* and *C.elegans*, thereby reduced likelihood for off-target error. Statistical analysis showed that minimizing off-target error and enhancing siRNA efficacy can be performed independently.

In summary, experimental RNAi targeted by siRNA has a certain degree of specificity. However, off-target effects yielding unintentional knockdown of unrelated genes are probable. The random occurrence of some level of sequence identity (including imperfect match) between siRNA and multiple targets in a transcriptome contributes to this undesired effect. The computational methods applied here may underestimate the off-target effects because of fairly stringent matching of sequence identity. Further studies will consider more relaxed rules for siRNA–target interaction such as bulge and wobble effects that occur *in vivo*. Although off-target effects can be reduced by minimizing sequence similarity with known transcripts and by rational design, it is recommended to include controls for specific targeting in RNAi experiments. Further understanding of siRNA will lead to more precise targeting of RNAi and reduce off-target effect to benefit the study of gene function and other future applications of RNAi.

ACKNOWLEDGEMENTS

The authors thank Vladimir Vuksan for implementing the web tools. This work was supported by NIH under grant number P20RR18754 from the Institutional Development Award Programme of the National Center for Research Resource. C.M.A. is supported by NIH grant RO1-AI052363. Funding to pay the Open Access publication charges for this article was provided by NIH grant number P20RR18754.

Conflict of interest statement. None declared.

REFERENCES

1. Fire,A., Xu,S.Q., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.
2. Fraser,A.J.G., Kamath,R.S., Zipperten,P., Campos,M.M., Sohrmann,M. and Ahringer,J. (2000) Functional genomic analysis of *C.elegans* chromosome I by systemic RNA interference. *Nature*, **408**, 325–330.
3. Dillin,A. (2003) The specifics of small interfering RNA specificity. *Proc. Natl Acad. Sci. USA*, **100**, 6289–6291.
4. Agrawal,N., Dasaradhi,P.V.N., Mohmmed,A., Malhotra,P., Bhatnagar,R.K. and Mukherjee,S.K. (2003) RNA interference: biology, mechanism and applications. *Microbiol. Mol. Biol. Rev.*, **67**, 657–685.
5. Jacque,J.M., Triques,K. and Stevenson,M. (2002) Modulation of HIV-1 replication by RNA interference. *Nature*, **418**, 435–438.
6. Surabhi,R. and Gaynor,R. (2002) RNA interference directed against viral and cellular targets inhibits human immunodeficiency virus type 1 replication. *J. Virol.*, **76**, 12963–12973.
7. Xia,H., Mao,Q., Eliason,S.L., Harper,S.Q., Martins,I.H., Orr,H.T., Paulson,H.L., Yang,L., Kotin,R.M. and Davidson,B.L. (2004) RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. *Nature Med.*, **10**, 816–820.
8. Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244–251.
9. Borkhardt,A. (2002) Blocking oncogenes in malignant cells by RNA interference—new hope for a highly specific cancer treatment? *Cancer Cell*, **2**, 167–168.
10. Barik,S. (2004) Development of gene-specific double-stranded RNA drugs. *Ann. Med.*, **36**, 540–551.
11. Check,E. (2004) Hopes rise for RNA therapy as mouse study hits target. *Nature*, **432**, 136.
12. Soutschek,J., Akinc,A., Bramlage,B., Charisse,K., Constien,R., Donoghue,M., Elbashir,S., Geick,A., Hadwiger,P., Harborth,J. *et al.* (2004) Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature*, **432**, 173–178.
13. Chi,J.T., Chang,H.Y., Wang,N.N., Chang,D.S., Dunthy,N. and Brown,P.O. (2003) Genomewide view of gene silencing by small interfering RNAs. *Proc. Natl Acad. Sci. USA*, **100**, 6343–6346.
14. Kamath,R.S., Fraser,A.G., Dong,Y., Poulin,G., Durbin,R., Gotta,M., Kanapin,A., Le Bot,N., Moreno,S., Sohrmann,M. *et al.* (2003) Systematic function analysis of the *C. elegans* genome using RNAi. *Nature*, **421**, 231–237.
15. Hsieh,A.C., Bo,R., Manola,J., Vazquez,F., Bare,O., Khvorova,A., Scaringe,S. and Sellers,W.R. (2004) A library of siRNA duplexes targeting the phosphoinositide3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic Acids Res.*, **32**, 893–901.
16. Catalanotto,C., Azzalin,G., Macino,G. and Cogoni,C. (2000) Transcription—gene silencing in worms and fungi. *Nature*, **404**, 245.
17. Ullu,E., Tschudi,C. and Chakraborty,T. (2004) RNA interference in protozoan parasites. *Cell. Microbiol.*, **6**, 509–519.
18. Tuschl,T., Zamore,P.D., Lehmann,R., Bartel,D.P. and Sharp,P.A. (1999) Targeted mRNA degradation by double-stranded RNA *in vitro*. *Genes Dev.*, **13**, 3191–3197.
19. Elbashir,S., 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.
20. Semizarov,D., Frost,L., Sarthy,A., Kroeger,P., Halbert,D.N. and Fesik,S.W. (2003) Specificity of short interfering RNA determined through gene expression signatures. *Proc. Natl Acad. Sci. USA*, **100**, 6347–6352.
21. Elbashir,S.M., Martinez,J., Patkaniowska,A., Lendeckel,W. and Tuschl,T. (2001) Functional atonamy of siRNA for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.*, **20**, 6877–6888.
22. Saxena,S., Jonsson,Z.O. and Dutta,A. (2003) Small RNAs with imperfect match to endogenous mRNA repress translation. *J. Biol. Chem.*, **278**, 44312–44319.
23. Jackson,A.L., Bartz,S.R., Schelter1,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. *Nat. Biotechnol.*, **21**, 635–637.
24. Scacheri,P.C., Rozenblatt-Rosen,O., Caplen,N.J., Wolfsberg,T.G., Umayam,L., Lee,J.C., Hughes,C.M., Shanmugam,K.S., Bhattacharjee,A., Meyerson,M. and Collins,F.S. (2004) Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells. *Proc. Natl Acad. Sci. USA*, **101**, 1892–1897.
25. 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.
26. Elbashir,S.M., Harborth,J., Weber,K. and Tuschl,T. (2002) Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods*, **26**, 199–213.
27. Reynolds,A., Leake,D., Boese,Q., Scaringe,S., Marshall,W.S. and Khovorova,A. (2004) Rational siRNA design for RNA interference. *Nat. Biotechnol.*, **22**, 326–330.
28. Hajarnavis,A., Korf,I. and Durbin,R. (2004) A probabilistic model of 3′ end formation in *Caenorhabditis elegans*. *Nucleic Acids Res.*, **32**, 3392–3399.
29. Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
30. Lai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.
31. Lewis,B.P., Shih,I.-H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
32. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.
33. Pancoska,P., Moravek,Z. and Moll,U.M. (2004) Effcient RNA interference depends on global context of the target sequence: quantitative analysis of silencing effciency using Eulerian graph representation of siRNA. *Nucleic Acids Res.*, **32**, 1469–1479.
34. Amarzguioui,M. and Prydz,H. (2004) An algorithm for selection of functional siRNA sequences. *Biochem. Biophys. Res. Commun.*, **316**, 1050–1058.
35. Chalk,A.M., Wahlestedt,C. and Sonnhammer,E.L. (2004) Improved and automated prediction of effective siRNA. *Biochem. Biophys. Res. Commun.*, **319**, 264–274.
36. Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W.S. (2003) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **1**, 1–10.
37. Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley, NY.
38. Qiu,S. and Lane,T. (2005) String kernels of imperfect matches for off-target detection in RNA interference. In Sunderam,V., Albada,G.D., Sloot,P.M.A. and Dongarra,J.J. (eds), Proceedings of Fifth International Conference on Computational Science, Lecture Notes in Computer Science. Springer-Verlag, Atlanta, GA (to appear).
39. Djikeng,A., Shi,H., Tschudi,C. and Ullu,E. (2001) RNA interference in *Trypanosoma brucei*: cloning of small interfering RNAs provides evidence for retroposon-derived 24–26-nucleotide RNAs. *RNA*, **7**, 1522–1530.
40. Horesh,Y., Amir,A., Michaeli,S. and Unger,R. (2003) A rapid method for detection of putative RNAi target genes in genomic data. *Bioinformatics*, **19** (Suppl. 2), ii73–ii80.
41. Amir,A., Landau,G., Keselman,D., Lewenstein,M., Lewenstein,N. and Rodeh,M. (2000) Text indexing and dictionary matching with one error. *J. Algorithms*, **37**, 309–325.
42. Garcia-Molina,H., Ullman,J.D. and Widom,J.D. (2002) *Database Systems: The Complete Book*. Prentice Hall Inc, NJ.
43. Khvorova,A., Reynolds,A. and Jayasena1,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209.
44. 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.
45. Sætrom,P. and Ola Snøve,J. (2004) A comparison of siRNA efficacy predictors. *Biochem. Biophys. Res. Commun.*, **321**, 247–253.
46. Yoshinari,K., Miyagishi,M. and Taira,K. (2004) Effects on RNAi of the tight structure, sequence and position of the targeted region. *Nucleic Acids Res.*, **32**, 691–699.
47. Luo,K.Q. and Chang,D.C. (2004) The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochem. Biophys. Res. Commun.*, **318**, 303–310.
48. Yang,D., Buchholz,F., Huang,Z., Goga,A., Chen,C.-Y., Brodsky,F.M. and Bishop,M.J. (2002) Short RNA duplexes produced by hydrolysis with *Escherichia coli* RNase III mediate effective RNA interference in mammalian cells. *Proc. Natl Acad. Sci. USA*, **99**, 9942–9947.
49. Nielsen,H., Engelbrecht,J., Brunak,S. and vonHeijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
50. 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.