Chapter 11

Designing Efficient and Specific Endoribonuclease-Prepared siRNAs

Vineeth Surendranath, Mirko Theis, Bianca H. Habermann, and Frank Buchholz

Abstract

RNA interference (RNAi) has grown to be one of the main techniques for loss-of-function studies, leading to the elucidation of biological function of genes in various cellular systems and model organisms. While for many invertebrates such as Drosophila melanogaster (D. melanogaster) and Caenorhabditis elegans (C. elegans) long double-stranded RNA (dsRNA) can directly be used to induce a RNAi response, chemically synthesized small interfering RNAs (siRNAs) are typically employed in mammalian cells to avoid an interferon-like response triggered by long dsRNA (Reynolds et al., RNA 12:988–993, 2006). However, siRNAs are expensive and beset with unintentional gene targeting effects (off-targets) confounding the analysis of results from such studies. We, and others, have developed an alternative technology for RNAi in mammalian cells, termed endoribonuclease-prepared siRNA (esiRNA), which is based on the enzymatic generation of siRNA pools by digestion of long dsRNAs with recombinant RNase III in vitro (Yang et al., Proc Natl Acad Sci USA 99: 9942–9947, 2002; Myers et al., Nat Biotechnol 21:324–328; 2003). This technology has proven to be cost-efficient and reliable. Furthermore, several studies have demonstrated that complex pools of siRNAs, as inherent in esiRNAs, which target one transcript reduce off-target effects (Myers et al., J RNAi Gene Silencing 2:181, 2006; Kittler et al., Nat Methods 4:337–344, 2007). Within this chapter we describe design criteria for the generation of target-optimized esiRNAs.

Key words: RNA interference, esiRNA, siRNA, Off-target effects, siRNA pool, siRNA efficiency

1. Introduction

RNA interference (RNAi) is a cellular process that takes part in the control of gene expression. The molecular machinery of RNAi is required for both the production and control of micro-RNAs (miRNAs), as well as small interfering RNAs (siRNAs). It is believed that RNAi has several functions in an organism or a cell, including the defense against parasitic genes from viruses or transposons (6), and the control of developmental processes through
posttranscriptional gene regulation (7). In the laboratory, RNAi has advanced to a powerful and widely used method to silence genes in cells or whole organisms (8).

Several enzymatic functions are part of the RNAi pathway. The endoribonuclease Dicer digests long double-stranded RNA (dsRNA) or miRNA precursors into small pieces of about 20–25 nucleotides, leading to the double-stranded form of siRNAs and miRNAs (9, 10). These RNAs activate the RISC (RNA induced silencing complex), which unwinds the si- or miRNAs, and incorporates one of the strands, the so-called guide strand (11). The RISC complex uses the enclosed guide strand to identify target mRNAs, which leads to their translational repression or degradation (11). Argonaute, the catalytically active RNase in the RISC complex, is involved in the selection of the RNA strand that is incorporated into the complex. It seems that the strand with the lower thermodynamic stability at its 5’end is preferentially chosen, suggesting that unwinding and strand incorporation are linked to each other (11). This step is a crucial factor in target specificity, as the RISC complex can only regulate the expression of target mRNAs complementary to the guide strand.

Most eukaryotic organisms, including mammals, express the RNA machinery to regulate their genes in a variety of biological processes, and this fact has been widely exploited to perform functional screens in invertebrate model organisms and plants. In these model systems, long dsRNA can be used for the silencing of genes. However, long dsRNA, when applied to most mammalian cells led to unspecific responses and cell death due to triggering an interferon-like response (1). In contrast, with the discovery that long dsRNA is processed in vivo and that RNAi is indeed mediated by short 21-mers (12), RNAi also became available for mammalian cells (13). This work has paved the way for making RNAi one of the most widely and successfully used techniques for functional genomics studies in mammalian cells. Yet it seemed that the usage of synthetic siRNAs also led to silencing of unintended targets (14), making results of gene knockdowns sometimes difficult to interpret. Strikingly, many of the affected unintended targets did not share similarity over the entire length of the siRNA, but rather resembled the miRNA-based silencing mechanism based on seed matches in the 3’UTR (15). Hence, the unspecificity of siRNAs probably results from eliciting a miRNA-like response (15). Since this discovery, much effort has been focused on improving the specificity of RNAi experiments in mammalian cells. Some progress has been reported by carefully selecting siRNA sequences (16) and/or by modifying certain bases in the RNA molecules (17). We and others could show that the off-target effect of the siRNAs could be greatly reduced by using complex mixtures of siRNAs targeting a gene rather than a single siRNA (4, 5) (Fig. 1).
Even before the increased specificity of siRNA mixtures was revealed, research laboratories explored alternatives to chemically synthesized siRNAs for usage in mammalian cells. A sensible approach was to take the step of dsRNA digestion out of the cell into the test tube and transfect cells with the resulting mixture of siRNAs (2, 18). Endoribonuclease-prepared siRNAs (esiRNAs) have since been successfully used to perform genome-scale knockdown screens (19–27). While the first screens were carried out using esiRNAs based on cDNA clones (28), later screens were based on target-optimized esiRNAs through in silico selection of regions most suitable for gene silencing (5).
Mechanistic insights into the RNA interference pathway (29, 30) and systematic studies of silencing efficiency and specificity (16, 31) have taught us about sequence features that should be considered when selecting an efficient siRNA. As stated earlier, the strand bias of the RISC complex for instance favors the incorporation of the strand with a thermodynamically less stable 5' end. Based on experimental and computational work, preferences for several positions in a siRNA have been defined (13, 16, 32–38), and systematic analysis of observed off-targets has led to rules governing the specificity of a siRNA (15, 39). When designing an esiRNA, the same principles apply, aiming for a mixture of optimized siRNAs with high potency and specificity. In this chapter, we introduce a step-by-step protocol (Fig. 2) to design specific and efficient esiRNAs for gene knockdown studies.
The design of esiRNAs involves two primary steps in succession. The first is the identification of the entity to be knocked down and the subsequent sequence region from which to design the esiRNA. The second is to predict, in silico, the silencing potential (the efficiency) and the off-target effect (the specificity) of the esiRNA.

mRNA sequences can be obtained from NCBI (http://www.ncbi.nlm.nih.gov) or ENSEMBL (http://www.ensembl.org) for most model organisms. When designing esiRNAs for use in *D. melanogaster* or *C. elegans*, sequences are preferably obtained from FlyBase (http://www.flybase.org) or WormBase (http://www.wormbase.org).

Because an esiRNA is designed against a specific sequence region, first, this region has to be determined. For efficient knockdown, the region against which the esiRNA is designed should be of a length ranging between 300 and 600 base pairs. This is to ensure efficient generation of dsRNAs and to target a region that does not code for highly conserved domains and has a high density of efficient constituent siRNAs. Furthermore, a minimum length of 300 base pairs ensures sufficient complexity of the resulting esiRNA pool for maximum specificity. As many genes are differentially spliced, the first step is to determine whether an esiRNA should target all possible splice variants or only a specific isoform.

In both cases, it is necessary to find a sequence common to all of the splice variants of a gene. In the former case, the esiRNA will be designed against this common region, while in the latter, the design process will consider sequence regions that are exclusive to the mRNA sequence to be targeted.
1. Look for the gene of interest at [http://www.ncbi.nlm.nih.gov/gene/advanced](http://www.ncbi.nlm.nih.gov/gene/advanced). Additionally, when querying NCBI at this page, in the Search Builder, choose the organism field, and specify the organism of interest. It is also possible to use ENSEMBL ([http://www.ensembl.org](http://www.ensembl.org)) to find sequences of interest, leveraging the BioMart utility available on the Web site. For the purposes of this protocol, we will adhere to NCBI. All of these steps can be easily applied to sequences retrieved from elsewhere as well. For designing esiRNA for a set of genes, sequences can be retrieved automatically (see Note 1).

2. On the relevant gene page, scroll down to the mRNA and protein(s) subsection of the NCBI Reference Sequences (RefSeq) section. Retrieve the FASTA sequences for each of the transcripts listed by choosing the FASTA (text) option from the Display Settings menu on the transcript page. If the gene has only one splice form, then this entire sequence will be used to design an esiRNA.

3. Use the procedure described under “Identification of the longest common substring (LCS)” in the Supplementary Methods section of (5) to determine a sequence stretch common to all of the transcripts. This procedure extends the suffix arrays idea of Manber and Meyers (40) by way of creating suffixes of multiple strings. An implementation of the procedure is available from the authors upon request.

4. If all of the splice forms of a gene are to be targeted, the region found in step 3 can be used for the esiRNA design. If a specific mRNA is to be targeted, then iteratively find the LCS between the mRNA of interest and all the other splice variants. Remove these regions from the specific target mRNA, and continue with the remaining sequence regions.

Over the course of the last decade, with the pervasive use of siRNAs as tools for loss-of-function studies, many parameters derived from experimental data have been evaluated with mixed results (13, 16, 32). These parameters address questions of thermodynamic properties, stability, and positional nucleotide preferences. It is prudent to use a set of parameters recurring across studies as predictors of siRNA efficiency.

With the sequence region determined as in Subheading 3.1, use a sliding window of 21 base pairs from the start of the sequence to enumerate the constituent siRNAs. Each of these siRNAs is to be checked for the sequence features described below. Assign a weight factor to each of the parameters under consideration; if a particular parameter does not have the desired value, this weight factor is added to the theoretical efficiency penalty (for suggested weight factors, see ref. 41). The objective of this procedure is to deduce a region whose sum of penalties, for the constituent siRNAs, is lowest.
1. The siRNA should be asymmetric, with an A/T at its 5’ end and a G/C at the 3’ end (33).
2. There should be no polynucleotide stretches of length more than 3 in the siRNA sequence (34, 35).
3. The GC content of the siRNA should be between 20% and 50% (16).
4. Positional nucleotide preferences: A at position 3 of the siRNA sequence, T at position 10, no G at position 13 and A or T at position 19 (16, 36–38).

siRNAs with near-perfect sequence similarity to an unintended target tend to down regulate that mRNA, a phenomenon known to induce off-targeting. Off-target effects produce confounding findings when used for loss of function studies (14). Subsequent studies interrogating off-target effects have found that siRNAs seem to behave like microRNAs, in that a hexamer or heptamer beginning at the second position in the siRNA sequence having matches in the 3’ UTR region of an mRNA tends to down regulate the mRNA (15). When predicting the specificity of the constituent siRNAs of an esiRNA, both of these modes of target down regulation have to be considered. Use the steps described below to analyze the 21 bp siRNA pool produced in Subheading 3.2.

While it has been a standard procedure to use BLAST to find short sequences with near-perfect matches, Bowtie (42), a program designed for analysis of Second Generation Sequencing data, is much faster, and its results are easier to process.

2. Use the bowtie-build program to build an index of the downloaded RefSeq FASTA file.
3. Run bowtie with the options -v 1 --norc --all iterating through the constituent siRNA sequences; the -v option specifies the number of mismatches allowed, the --norc tells bowtie to map only in the forward orientation.
4. For each constituent siRNA sequence, remove the match/es corresponding to the intended target/s, and count the remaining mRNAs that have near-perfect complementarity to the siRNA sequence. A region with high numbers of near-perfect matches should not be considered for further esiRNA production.

2. Locate the 3’UTR of the transcripts. For this, use for instance the BioPython framework (or one of the other frameworks corresponding to the programming language of choice—listed at http://www.open-bio.org) to parse the downloaded GenBank file to determine the end of the CDS.

3. For each constituent siRNA sequence, search the 3’UTR sequences corresponding to mRNAs other than the intended mRNA target/s for matches with the hexamer and heptamer starting at position 2 in the siRNA sequence; count the number of occurrences of the hexamer and heptamer in the 3’UTR sequences. If esiRNAs are to be designed frequently, it would be advisable to create a repository of seeds (see Note 2).

4. If there are three or more matches to the constituent siRNA sequence’s seed region, then flag the mRNA corresponding to the 3’UTR in which the matches are found as an off-target. Birmingham and colleagues (15) concluded from their analysis of experimental data that three or more matches to the seed region have 100% predictive power for determining miRNA-like targets.

The penalties assigned in Subheadings 3.2 and 3.3 corresponding to the violation of efficiency conditions and the off-targeting pervasiveness have to be combined to pick a region that has the minimal penalty across its constituent siRNAs, such that this fragment has the highest density of efficient and specific silencers.

1. Initialize a 1-dimensional vector with size equal to the length of the sequence region determined in Subheading 3.1.

2. For each index of the 1 dimensional vector, find all the siRNAs constituting an esiRNA of a range of lengths from a minimum of 300 to a maximum of 600 starting at that same index in the sequence region determined in Subheading 3.1. Note that each element of the 1 dimensional vector will now be a set of sets of siRNAs.

3. For each set of siRNAs in every element of the vector, find the average efficiency penalty and the average off-target count corresponding to the siRNAs as determined in Subheadings 3.2 and 3.3.

4. For every element of the vector, corresponding to each esiRNA in that element, compute a rank based on the average efficiency penalty and the average off-target count. Compute the average rank for each esiRNA in that element, and assign
to the element the average efficiency penalty and average off-target count corresponding to the esiRNA with the least average rank.

5. Repeat the ranking procedure described in step 4, but now applying it to the entire vector.

6. Find the element of the vector with the least average rank. The position of the minimal element yields the starting position of the most optimal stretch in the sequence region that can be endoribonuclease digested resulting in a pool of silencing triggers.

Given that these steps are quite elaborate, to design esiRNAs frequently or for large sets of genes, it would be prudent to automate the procedure (see Note 3). On the other hand, for knockdown of a single gene, it would be easier to use a Web service that already implements the analysis described in this chapter (see Note 4).

4. Notes

1. For designing esiRNAs for a set of genes, it is advisable to retrieve the sequences using the E-utilities framework from NCBI (http://www.ncbi.nlm.nih.gov/books/NBK25500) or the ID filter in BioMart at ENSEMBL (http://www.ensembl.org/biomart/martview).

2. If the design of esiRNAs is a frequent task, or for a large set of genes, whilst checking for seed matches in the 3' UTR, it would be more efficient to preprocess the 3’ UTRs. This can be done by creating an index of all the seeds (both hexamers and heptamers) from the entire set of 3’ UTRs and querying this index.

3. For regularly designing esiRNAs, an object oriented framework implementing the design steps listed in this chapter would be ideal. Such a framework would afford a platform to easily incorporate changes in design principles accruing from continuing research into RNAi mechanisms.

4. To design single esiRNAs, a Web application such as DEQOR (http://deqor.mpi-cbg.de) can be used. An input sequence is scored based on the steps described in this chapter, and a graphical display (Fig. 3) eases the selection of a region from which to design an esiRNA.
Fig. 3. Screenshot of efficiency and specificity analysis of a coding sequence. The top panel indicates a graphical representation of individual siRNAs color coded for their efficiency and specificity: green bars represent siRNAs which are predicted to be very efficient with no predicted off-targets, black bars represent those which are not efficient and with no off-target and the red bars represent siRNAs with off-targets. The middle panel shows the sequence of the region that can be highlighted in the top-panel, while the bottom panel lists individual siRNAs and their predicted off-targets.

References


