E-RNAi: a web application to design optimized RNAi constructs

Zeynep Arziman, Thomas Horn and Michael Boutros*

Boveri-Group Signaling and Functional Genomics, German Cancer Research Center, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany

Received February 14, 2005; Revised and Accepted April 13, 2005

ABSTRACT
RNA interference (RNAi) has become a powerful genetic approach to systematically dissect gene function on a genome-wide scale. Owing to the penetrance and efficiency of RNAi in invertebrates, model organisms such as Drosophila melanogaster and Caenorhabditis elegans have contributed significantly to the identification of novel components of diverse biological pathways, ranging from early development to fat storage and aging. For the correct assessment of phenotypes, a key issue remains the stringent quality control of long double-stranded RNAs (dsRNA) to calculate potential off-target effects that may obscure the phenotypic data. We here describe a web-based tool to evaluate and design optimized dsRNA constructs. Moreover, the application also gives access to published predesigned dsRNAs. The E-RNAi web application is available at http://e-rnaire.dzkf.de/.

INTRODUCTION
Until recently, systematic reverse genetic approaches to probe loss-of-function phenotypes have been difficult to conduct. This has been largely due to the limitations in the ability to generate genome-wide collections of directed knock-out mutants. However, the discovery of post-transcriptional silencing mechanisms by small interfering RNAs (siRNAs) has allowed the development of tools to efficiently knock down expression of specific genes. First discovered in Caenorhabditis elegans, double-stranded RNA (dsRNA) molecules have been shown in subsequent studies to represent both important endogenous regulators of gene expression and a powerful new tool to silence gene expression (1,2). In C.elegans, several genome-scale collections of dsRNAs have been applied to study developmental defects, fat storage and other phenotypes (3–5), which has been facilitated by the ease of dsRNA introduction: feeding of worms with E.coli which express dsRNA molecules is sufficient to produce loss-of-function phenotypes. Although feeding dsRNA to Drosophila is not effective, the Drosophila system is tractable to RNAi in both embryos and cultured cells (6). Typically 300–700 bp of dsRNA are synthesized from DNA templates containing terminal T7 promoters and the resulting molecules are simply added to the culture medium. These dsRNA molecules are taken up by cells through an unknown transport mechanism and are intracellularly processed into functional 21mer siRNAs. Several well-characterized cell lines of embryonic origin are available and have been used to perform genome-scale RNAi screens for various phenotypes (7–13).

The efficiency of RNAi by long dsRNAs in invertebrates is probably due to the ‘natural’ pooling of many 21 nt sequences. Long dsRNAs are intracellularly cleaved into 21–22 nt siRNAs by the Dicer complex and direct the degradation of target mRNAs through the RNA-induced silencing complex (RISC) (2). Although a dsRNA should be designed to match to one specific gene, off-target effects can occur if siRNAs have sequence homology to genes that are not supposed to be targeted. Furthermore, the knock-down of target transcripts might differ depending on the efficiency of siRNAs derived from long dsRNAs. It has been proposed that parameters for siRNA efficiency include GC content, asymmetry and thermodynamic stability (14). In efficient siRNAs, the 5’ end of the anti-sense strand and the target site have relatively low thermodynamic stability whereas the 5’ end of the sense strand has a high thermodynamic stability. These thermodynamic properties appear to be important for promoting the incorporation of the anti-sense strand into the RISC, blocking the incorporation of the sense strand and promoting RISC-anti-sense strand mediated mRNA degradation (15–17). In order to design efficient and specific siRNAs for experiments in mammalian cells, a number of computational tools have been developed that incorporate recent design rules (18–20).

We have developed the E-RNAi web application to design and evaluate dsRNA constructs suitable for RNAi experiments in Drosophila and C.elegans. It can also be used for the design of enzymatically digested long dsRNA (esiRNAs) for mammalian cells (21). dsRNA sequences (RNAi probes)
are evaluated for their predicted specificity and efficiency. Since DNA templates used to generate dsRNAs are generated by PCR, primer pairs suitable to amplify DNA templates from genomic DNA or cDNA are calculated. In addition E-RNAi allows access to predesigned dsRNAs from published experiments.

**WEB APPLICATION**

Our aim was to create a web application to automate the design of optimized dsRNA constructs that are commonly used for RNAi experiments in invertebrate model organisms such as *Drosophila* and *C.elegans*. To this end, the E-RNAi application has to accomplish several tasks, including (i) the identification of the targeted transcript, (ii) *in silico* dicing of the template sequence into all possible siRNAs, (iii) calculation of the RNAi efficiency of each *in silico* diced siRNA, (iv) identification of sequences (siRNAs) that potentially target additional genes and (v) design of optimized PCR primers to amplify dsRNA templates directly from genomic DNA or cDNA. When pre-designed dsRNAs from publicly available RNAi libraries are available, the web application retrieves the information from a relational database. Although our web application mainly aims to create optimal probes for RNAi experiments in *Drosophila* and *C.elegans*, it can also be used to create long dsRNAs that can be ‘diced’ *in vitro* to be used in mammalian cells. A general outline of the program is shown in Figure 1.

**User input**

Three different run-time options are available in E-RNAi. The user can choose to design RNAi probes *de novo*, to retrieve predesigned probes or to evaluate an input sequence for its RNAi specificity and efficiency using transcriptome databases from *Drosophila*, *C.elegans* or human. The user can also deselect database and off-target evaluation if genomic and transcript information is not available. Other options for the *de novo* design include siRNA length for *in silico* dicing (default 21 bp) and primer design (primer size and primer product size). The number of primer pairs to be designed is crucial for probe optimization. A higher number of designed primer pairs increases the probability of identifying an optimal probe for a specific gene but comes at a cost in terms of computing time. The ‘probe retrieval’ option uses nucleotide as well as amino acid sequences as input and retrieves pre-designed dsRNAs from publicly available RNAi libraries of *Drosophila* and *C.elegans*. The ‘probe evaluation’ option runs a specificity and efficiency evaluation using the sequence information entered (Figure 2).

**Identification of primary targets**

First the sequence input is mapped to predicted transcripts using BLASTN against all predicted transcripts of the organism chosen by the user. This ‘primary’ target transcript is necessary to define which other genes are hit by the designed long dsRNA and to identify predesigned dsRNAs that are available in public libraries.

**In silico dicing**

Using a user-defined length, the program cuts the sequence into 18–25 nt long siRNAs with a 1 nt shifting window.

---

**Figure 1.** Schematic representation of the program. The web application was implemented as Perl scripts and CGI modules that relied on BioPerl 1.5 (26). The data are in parts, stored in a relational MySQL database. Sequence homology searches are performed using BLAST (27). Primers to amplify PCR templates for dsRNAs are automatically designed using a local implementation of Primer3 (28). Genes and RNA templates are visualized in their genomic context using the Generic Genome Browser (29). Several data sources were integrated and are accessible through E-RNAi, including transcript and gene data obtained from Flybase [Drosophila, (30)] and Wormbase [C.elegans, (31)] and predesigned RNAi libraries (9,24). RNAi phenotypes were collected from Flybase, Wormbase and recent publications and assembled into a relational database that allows cross-species phenotype comparisons (unpublished data).

In addition to the procedure shown here, the user can choose the option ‘probe retrieval’ to retrieve predesigned dsRNAs. When the ‘probe evaluation’ option is selected, the full-length input sequence is evaluated for efficiency and specificity parameters.

The specificity and efficiency of predicted siRNAs are then calculated for these *in silico* diced sequences.

**Calculation of siRNA efficiency**

The program predicts siRNA efficiency using an algorithm described by Reynolds et al. (22). The siRNAs are evaluated for GC content, low stability at the sense strand 3’-terminus, inverted repeats and base preferences (at positions 3, 10, 13 and 19 of the sense strand). The implemented algorithm estimates the efficiency of siRNAs using eight criteria: (i) low GC content (30–52%), (ii) at least three A/U bases at positions 15–19, (iii) absence of internal repeats, (iv) an A base at position 19, (v) an A base at position 3, (vi) a U base at position 10, (vii) a base other than G or C at position 19 and (viii) a base other than G at position 13. If an siRNA fulfills criteria (i), (ii), (iii), (iv) and (v), one point is added to its score. For a failure to fulfill criteria (ii) and (viii), one point is subtracted from the score. For criterion (i), one point is added for each A or U base in positions 15–19, up to a maximum of five points. For criteria (iv), potential hairpin structures of siRNAs are calculated using RNAfold (23). If the melting temperature of the potential hairpin region is 20°C or less, one point is added to the score. A siRNA with an efficiency score of six or higher is
In silico BLAST searches of E-RNAi predicts the specificity of a dsRNA by performing calculation of siRNA specificity of efficient siRNAs in a dsRNA. The percentage efficiency of a probe is calculated as the percentage considered an efficient silencer (22). In the output, the percentage efficiency of a probe is calculated as the percentage of efficient siRNAs in a dsRNA.

**Calculation of siRNA specificity**

E-RNAi predicts the specificity of a dsRNA by performing BLAST searches of in silico diced siRNAs against the selected transcriptome using a penalty for nucleotide mismatch of −3. BLAST searches are performed with standard settings without sequence filtering. The specificity score for an in silico diced siRNA is defined as one divided by the number of genes that are hit by a specific siRNA, whereby a hit is defined as a perfect match to a gene sequence. The complete dsRNA is then scored for an average percentage specificity for all possible siRNAs. The specificity of an RNAi construct for a certain gene and its alternative splice variants is calculated as the number of matching siRNAs over the number of all siRNAs in the dsRNA of interest.

**Display and export of results**

Figure 3 shows a typical result for RNAi probe de novo design. In the first table, the de novo designed RNAi probes are sorted by their overall score. This score is calculated by ranking their primer quality, their efficiency and their specificity. The weighted sum of primer quality, specificity and efficiency rankings leads to the final score and sorting of probes. Figure 3B shows an example of the RNAi probes retrieved from RNAi libraries which also target the queried gene.

**RNAi libraries**

The current database (version 1.0) contains RNAi probes from libraries that were designed to cover almost all open reading frames (ORFs) in the genomes of *Drosophila* and *C. elegans*. The Heidelberg/Boston RNAi library contains 21,300 dsRNAs that target almost all predicted genes in the *Drosophila* genome (9,24). About 13,100 probes are available in the MRC/Cyclacel library, which was designed on the basis of the Berkeley *Drosophila* Genome Project (BDGP) annotations, and cover ~90% of the *Drosophila* genome. Presdesigned probes can also be retrieved from an RNAi library of 18,041 dsRNAs covering 87% of the *C. elegans* genome (5,25,32). Additional libraries will be added as they become publicly available.

**EVALUATION OF LARGE-SCALE LIBRARIES**

Genome-wide RNAi libraries have been successfully used in *C. elegans* and *Drosophila* to systematically identify components of various cellular pathways. In order to benchmark individual dsRNA constructs that are part of genome-wide RNAi libraries, we applied the approach implemented in the E-RNAi software to evaluate RNAi libraries for both their predicted efficiency and their predicted specificity. Figure 4 shows the results calculated for the Heidelberg/Boston RNAi library, which targets almost every gene in the *Drosophila* genome (9,24). The RNAi probes contained in the library are homologous to a total of 12,929 genes in the *Drosophila* genome (25). In total, 5,760,284 siRNAs were computationally generated by in silico dicing using a length window of 21 nt. For each individual siRNA, we calculated efficiency and specificity scores according to the algorithms outlined above. The efficiency scores of the predicted siRNAs follow a normal distribution, with a mean efficiency score of 3.6 and a standard deviation of 2.0. In total, 18.8% of all siRNAs obtained a score ≤6, indicating that they are predicted to be efficient silencers. We then calculated the percentage of efficient siRNAs (with a score ≥6) per dsRNA construct. The distribution in Figure 4C shows that 5682 out of 14,300 dsRNAs contain 20% or more efficient siRNAs, whereas 3.5% have <5% predicted efficient siRNAs. Overall, the analysis shows that an RNAi probe with an average size of 402 bp contains 75 siRNAs that are predicted to be efficient silencers.

We then evaluated the specificity of all dsRNAs contained in the library by BLAST analysis of each individual siRNA against the *Drosophila* transcriptome. Of the 5,641,589 evaluated siRNAs, 101,203 (1.8%) showed more than one hit in the transcriptome. In total, 2715 dsRNAs contained at least one siRNA that potentially targets an unintended transcript. We then assessed how many off-target siRNAs can be considered to be efficient. About 84% of all cross-specific siRNAs were inefficient in silencing genes. Combining both efficiency and specificity analysis allows the conclusion that up to...
1083 (7.5%) of all dsRNAs contained in the Heidelberg/Boston RNAi library could have off-target effects.

The analysis of RNAi libraries showed that a large majority of RNAi probes are predicted to be specific, and ‘natural pools’ of diced dsRNA give rise to many efficient siRNAs that are likely to be sufficient to knock down the target transcript. Predicted problematic probes can be computationally ‘flagged’ in the downstream analysis of phenotypes.

Figure 3. Output of the E-RNAi software. (A) De novo designed RNAi probes against the Rel gene are sorted by their overall score. In order to calculate this score, primer quality is weighted with 0.2, specificity with 0.25 and efficiency with 1. The table contains the length of each probe, the specificity and efficiency of each probe and the quality of primer pairs. (B) RNAi probes retrieved from RNAi libraries designed against the query are listed and evaluated. (C) Detailed information about each de novo designed probe is provided in the output page. A primer sequence with the SP6 promoter tag sequence (written in small letters), primer properties and probe sequence are shown. (D) De novo designed and retrieved probes are displayed on a GBrowse image at the bottom of the output page.
Examples: *Rel* and *Mask* genes

To demonstrate the functionality of the E-RNAi web application for *de novo* design approaches, computational predictions of dsRNA templates for two transcripts were generated. The two transcripts were chosen because of their divergent properties. *Rel* is a transcription factor important for innate immune responses and is encoded by a compact gene in the *Drosophila* genome (Figure 5A). In contrast, *Mask* shows significant...
homology on both the nucleotide and the protein level to other Ankyrin-domain containing genes (Figure 5D). As input, E-RNAi was given the complete ORF of Rel and Mask for the de novo design of RNAi probes. Figure 5 shows the analysis of both transcripts for efficiency and specificity scores. Approximately 22.9% of the possible siRNAs against the Rel gene show an efficiency score of 6 or higher (Figure 5B). Specific dsRNAs can be designed for most regions of the Rel gene which contains only four unspecific siRNAs, whereas a dsRNA targeting the complete ORF of the Mask gene would contain 192 unspecific siRNAs (Figure 5C and F). About 65% of unspecific siRNAs hit more than 2 genes, and 40% hit more than 10 genes. About 19.2% of the Mask transcripts are suitable for efficient gene silencing (Figure 5E). Those inefficient and unspecific regions are dispersed along the gene, which can be problematic for the design of RNAi probes.

The analysis showed that some gene regions are more prone to yield unspecific or inefficient RNAi probes than others. A dsRNA against a gene such as Mask can result in unintended gene silencing because of stretches with high sequence homology. Therefore predicted efficiency and specificity scores should play an important role during dsRNA design in order to achieve high efficiency and minimum off-target effects.

DISCUSSION

Evaluation and de novo design of long dsRNAs for efficiency and specificity remains an important issue in the design of large-scale RNAi experiments both in model organisms and in human cells. The design of long dsRNAs has to take into account both restrictions based on the properties of the contained siRNAs, such as their specificity and efficiency, and experimental limitations, such as the identification of primer pairs that are necessary to amplify the dsRNA template sequence by PCR from cDNA or genomic sources. Here we present a web application that automates the required tasks, from the prediction of efficient and specific target sites to the design of appropriate primer sequences. Results for a specified number of calculated dsRNAs are presented in their genomic context and the user can choose to export sequences as a tab-delimited file.

Limitations in the prediction of siRNA efficiency and specificity remain. The algorithms to calculate siRNA efficiency will probably be improved in the future as more experimental evidence to predict siRNA efficiency becomes available. Similarly, the calculation of potential off-target effects might be dependent on which specific regions of siRNAs are sufficient for silencing, and improved search algorithms for relevant sequence homologies will probably improve prediction of off-target effects. It is our goal to continuously update the E-RNAi web application to include improved algorithms.

In addition to de novo design, the systematic analysis of already available dsRNAs remains an important issue. Large-scale RNAi libraries have been generated that target almost the complete transcriptome in Drosophila and C. elegans. A prediction of off-target effects and the efficiency of dsRNAs is important to assess both false positive and false negative rates in screening experiments. In particular, the systematic analysis of both positive and negative results from genome-wide RNAi experiments should benefit from the exclusion of transcripts that are targeted by off-target siRNAs or inefficient RNAi constructs.

ACKNOWLEDGEMENTS

We are grateful to Renato Paro and Marc Hild for providing sequence information for newly predicted genes. The MRC/Cyclacel Drosophila RNAi library was obtained through MRC geneservices (Cambridge, UK). We are also grateful to David Emmert and William Gelbart for help with Flybase linkouts. We would like to thank Kerstin Bartscherer and Florian Fuchs for comments on the manuscript and members of the Boutros Laboratory for helpful suggestions and discussions. We are grateful to Tobias Reber for support in computational infrastructure. The research was funded in part by a Grant in the Emmy-Noether Program of the German Research Foundation to M.B. Funding to pay the Open Access publication charges for this article was provided by intramural funds.

Conflict of interest statement. None declared.

REFERENCES


