Highly specific and accurate selection of siRNAs for high-throughput functional assays

Javier Santoyo, Juan M. Vaquerizas and Joaquín Dopazo*

Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO) and Functional Genomics Node, INB, Melchor Fernández Almagro, 3, 28029 Madrid, Spain

Received on August 12, 2004; revised and accepted on November 26, 2004
Advance Access publication December 10, 2004

ABSTRACT

Motivation: Small interfering RNA (siRNA) is widely used in functional genomics to silence genes by decreasing their expression to study the resulting phenotypes. The possibility of performing large-scale functional assays by gene silencing accentuates the necessity of a software capable of the high-throughput design of highly specific siRNA. The main objective sought was the design of a large number of siRNAs with appropriate thermodynamic properties and, especially, high specificity. Since all the available procedures require, to some extent, manual processing of the results to guarantee specific results, specificity constitutes to date, the major obstacle to the complete automation of all the steps necessary for the selection of optimal candidate siRNAs.

Result: Here, we present a program that for the first time completely automates the search for siRNAs. In SIDE, the most complete set of rules for the selection of siRNA candidates (including G+C content, nucleotides at determined positions, thermodynamic properties, propensity to form internal hairpins, etc.) is implemented and moreover, specificity is achieved by a conceptually new method. After selecting possible siRNA candidates with the optimal functional properties, putative unspecific matches, which can cause cross-hybridization, are checked in databases containing a unique entry for each gene. These truly non-redundant databases are constructed from the genome annotations (Ensembl). Also intron/exon boundaries, presence of polymorphisms (single nucleotide polymorphisms) specificity for either gene or transcript, and other features can be selected to be considered in the design of siRNAs.

Availability: The program is available as a web server at http://side.bioinfo.cnio.es. The program was written under the GPL license.

Contact: jdopazo@cnio.es

INTRODUCTION

RNA interference (RNAi) is a sequence-specific posttranscriptional gene-silencing process that constitutes a powerful new tool for analyzing gene knockdown phenotypes (Fire et al., 1998). This mechanism is evolutionary conserved among eukaryotes and plays an essential role in mediating responses to exogenous RNAs (such as viruses) and in stabilizing the genome by sequestering repetitive sequences (such as transposons; see Hannon (2002) and Tijsterman et al. (2002)). RNA silencing is triggered by double-stranded RNA molecules, which are cleaved by the enzyme DICER into 21–23 nt duplexes containing a 2 nt overhang at the 3′ end of each strand. These duplexes are incorporated into a protein complex called the RNA-induced silencing complex (RISC). The RISC is responsible for the sequence-specific degradation of target RNAs that contain homologous sequences. The resulting down-modulation of the encoded proteins can subsequently lead to the induction of a specific phenotype.

Gene-specific long dsRNAs have been successfully used in Caenorhabditis elegans and Drosophila melanogaster for RNAi-mediated gene silencing. In mammalian cells, however, dsRNAs longer than 30 nt trigger the antiviral/interferon pathways, which results in global shutdown of protein synthesis (de Haro et al., 1996). It was recently demonstrated that RNAi-mediated gene silencing can be obtained in cultured mammalian cells by delivery of either chemically synthesized short (21–23 nt) double-stranded siRNA molecules or by endogenous expression of short hairpin RNAs bearing fold-back stem–loop structure. These siRNAs are long enough to induce gene-specific suppression, although short enough to evade the host interferon response (Elbashir et al., 2001a).

The effectiveness of a siRNA is likely to be determined by the accessibility of its target sequence in the intended substrate. Many of the siRNAs reported to date are designed to target coding sequences. There are no reliable ways, however, of predicting or identifying the ‘ideal’ sequence for a siRNA and the selection of siRNAs sequences are largely empirical. General guidelines for designing siRNA oligonucleotides have been proposed which require avoiding regions of mRNA, such as 5′-untranslated region (5′-UTR) and 3′-UTR and regions within 75 bases of a start codon and sequences having no extreme G+C contents. Recent works have allowed the identification of a number of siRNA-specific features with a significant contribution to the different steps of the RNAi-mediated gene silencing process and different authors proposed sets of more accurate rules. These new rules, experimentally validated with a large set of siRNAs and target genes, comprise different requirements on highly specific determinants of siRNA functionality. Thus, Reynolds et al. (2004) proposed a score obtained as a weighted summation of a set of parameters including G+C content (ranging between 30 and 52%), presence of determined nucleotides at specific positions and propensity to form internal hairpin structures is highly correlated with the probability of a candidate siRNA acting as a potent silencer. Amarzguioui and Prydz (2004) identify features consistently correlated with functionality that include mainly an asymmetry in the stability of the duplex ends and the motifs S1, A6 and W19, which are used for the score. Ui-Tei et al. (2004) empirically found that highly effective RNAi
must satisfy the following sequence conditions at the same time: (1) A/U at the 5′ end of the antisense sequence; (2) G/C at the 5′ end of the sense sequence; (3) AU-richness in the 5′ terminal, 7 bp long region of the AS; and (4) the absence of any long GC stretch of >9 bp in length. Takasaki et al. (2004) propose a selection method based on the gene degradation measure (priority score) defined by positional features of individual nucleotides. Finally, Hsieh et al. (2004) identify different sequence features as being important for an efficient silencing that can be used to score siRNA candidates.

Furthermore, to ensure siRNAs work properly in gene knockdown experiments, siRNA-mediated transcriptional silencing must be specific. The possibility of cross or co-suppression between closely related genes or by the presence of a highly conserved region within the gene is a major potential cause of artifacts. Consequently, a rigorous selection process for RNAi targets ideally includes the relatively simple but laborious process of similarity searches against available data for the organism under study.

With the advent of large-scale genome sequencing projects, thousands of genes whose functions remain unknown have been identified. Reverse genetics is the most effective way to assess the function of a gene, but to date there has been no general method for reverse genetics other than gene targeting by homologous recombination, which is both time consuming and costly. Currently, RNAi-mediated gene silencing has become a reliable method for the study of gene function in a genome-scale context (Asharaf et al., 2003). However, not all possible siRNA duplexes targeting a gene are equally effective, hampering the development of large-scale projects. In order to design siRNAs for a large number of genes, automatic tools for pattern and specificity selection need to be implemented. Several tools have been developed for finding siRNA candidates. Among them, Sirna (http://sifold.wadsworth.org/sirna.pl) is a specialized tool for target accessibility prediction and RNA duplex thermodynamics for rational siRNA design based on an advanced thermodynamics algorithm (Ding and Lawrence, 2003). The EMBOSs siRNA (http:// Athena.bioc.uvic.ca/cgi-bin/emboss.pl?_action=inputs&_app=sirna) incorporates rules defined by several authors (Elbashir et al., 2001a,b) but does not perform any specificity check. SiSearch (Chalk et al., 2004), by Sonnhammer group, uses a series of pattern matching rules and some thermodynamic rules as described by Mathews et al. (1999). These rules generate a set of rules with a certain overlap and use the Unigene database for specificity checking. The Oligo Retriever from the Cold Spring Harbour Laboratories (http://katabin.cshl.org:9331/RNAi/html/maІ.html) is based on the rules used by Paddison et al. (2002) for the RNA-mediated gene silencing. Recently, gene-specific siRNA selector (Levenkova et al., 2004) and the siRNA selection server at Whitehead Institute (Yuan et al., 2004) have been released with similar pattern matching and energy rules but with a more careful pipeline for selecting specific siRNAs. Nevertheless, as highlighted in the program guidelines of the first one, the use of Unigene and Locuslink databases for checking specificity does not completely guarantee specificity in the siRNA targets found and, what is more important, hamper the possibility of an automatic processing of the results because human intervention is required. In a different approach, DEQOR (Henschel et al., 2004) uses genomic sequences together with transcriptomes to check for cross-specificity, although the authors recommend a careful examination of the results in the case of large number of genes is under screening for siRNAs candidates.

Trying to overcome the specificity problem in large-scale siRNA selection other approaches have been proposed. Recently, siDirect (Naito et al., 2004) uses a database constructed by aligning the human RefSeq and Unique Unigene databases onto the genomic sequences (and the same is done for mouse sequences). The process is slow and cumbersome, requires continuous updates and do not guarantee true non-redundancy.

Moreover, none of the above-mentioned programs makes complete use of the new heuristic scores proposed by Reynolds et al. (2004), which have been proven to suggest siRNA candidates with higher likelihood of producing efficient silencing. Only the Dharmacon siDESIGN Center (http://design.dharmacon.com/madesign/default.aspx), to our knowledge, includes these rules. Nevertheless, this program, as well as the other above-mentioned programs, can only process one sequence at a time, which do not make them suitable for high-throughput design.

Here, we describe a new tool for the automatic design of human and mouse siRNA, SiDE (which stands for Small interfering RNA design), that applies different rules for pattern sequence selection, including the most accepted Reynolds et al. (2004) heuristic scores, and performs similarity searches in order to minimize cross-hybridization. In addition, the presence of genomic features such as single nucleotide polymorphisms (SNPs) can be considered in the design.

The tool is designed for high-throughput design of siRNAs, which implies high specificity and capacity for dealing with a large number of target genes. In SiDE, a conceptually new way of checking the specificity of siRNAs has been implemented by using real non-redundant databases obtained from the genomes annotated in Ensembl (Birney et al., 2004). To make the application as widely available as possible, a web-based interface to the program has been developed (http://side.bioinfo.cnio.es).

**METHODS**

The main objective of the tool is the high-throughput design of siRNAs. To achieve this, we have developed a program capable of processing a large number of genes and producing lists of highly specific, putative siRNAs that, at the same time, fulfill a series of thermodynamic and empirical rules. SiDE includes state-of-the-art rules for the design of functional siRNAs including the rules based on the empirical work recently published by different authors (Reynolds et al., 2004; U-I Tei et al., 2004; Amarzguioui and Prydz, 2004; Hsieh et al., 2004; Takasaki et al., 2004), which have proven to be more efficient than the consensus rules accepted to date. The novelty of SiDE is the high specificity in the detection of targets for siRNA interference. Currently, SiDE can find siRNA candidates in humans and mice.

**siRNA candidate selection**

The starting point is a list of genes that can be provided using different identifiers that currently include HUGO systematic names, RefSeq and locusLink IDs, Ensembl IDs, and EMBL, TEmBL, SwissProt and Uniprot gene identifiers (AC codes). Sequences of all the transcripts annotated for each gene in the list are retrieved from the Ensembl database (Birney et al., 2004). Each transcript sequence is scanned for candidate siRNA sequences that comply with a set of rules. Most of these rules are semi-empirical and are based on the work of different authors (Elbashir et al., 2001a,b; Jackson et al., 2003; Semizarov et al., 2003; Reynolds et al., 2004) and implemented as parameters that can be modified (within certain limits) by the user. All the candidate siRNA sequences can be checked for gene-specificity using the BLAST program (Altschul et al., 1990) on a non-redundant database that contains unique copies of each gene, generated as explained below. The first hit must be the target sequence. If the next hit is below a threshold of similarity (which indicates a low likelihood of cross-hybridization), it is accepted and if not, then discarded.
Parameters for siRNA design

The set of parameters that can be modified by the user allows a completely customized design of siRNAs based either on the ‘classical’ parameters, or in different score schemes, obtained from empirical studies, recently proposed by distinct authors, or in any desirable combination. Figure 1 shows a snapshot of the web interface with all the available parameters. The design parameters are as follows:

(a) **ORF target** Define limits for searching candidate siRNAs with respect to the start and stop codons. It is advisable to avoid 5’- and 3’-UTRs as well as the neighbourhood of the start and stop codons because these regions are rich in regulatory motifs and UTR-binding proteins and translation initiation complexes could interfere with siRNA binding. Within the defined limits (75 nt after the start codon and 50 nt before the stop codon, by default), a sliding window, typically of 23 bp (although patterns of different length can also be used) is used to scan downstream for siRNAs candidates. All candidates must meet the additional parameters as listed below.

(b) **Patterns**. Candidate siRNAs can be searched to fit one or more predefined patterns (AAN19TT, NARN17YNN, NNN19NN, AAN19NN, NANN17YNN, AAN19, NAN19NN, AARN18NN, NNN19, being letters from IUPAC codes: N any nucleotide, R purines and Y pyrimidines). Selection of more degenerate patterns discard the use of more specific ones (e.g. selection of AAN19NN will disable the pattern AAN19TT). Furthermore, user-defined patterns covering between 19 and 23 nt can be used. The rationale for these patterns is to generate a symmetric duplex with respect to the sequence composition of the sense and antisense 3’ overhangs (see Elbashir et al., 2001a,b and Tuschl website, http://www.rockefeller.edu/labheads/tuschl/sirna.html). The first suggested choice is searching for 23 nt sequence motif AA(N19)TT and select hits with ∼50% G/C-content (30–70% has also worked in for them). If no suitable sequences are found, the search is extended using the motif NA(N21). The sequence of the sense siRNA corresponds to (N19)TT or N21 (position 3–23 of the 23 nt motif), respectively. The antisense siRNA is synthesized as the complement to position 1–21 of the 23 nt motif. Given that position 1 of the 23 nt motif is not recognized sequence-specifically by the antisense siRNA, the 3’-most nucleotide residue of the antisense siRNA, can be selected deliberately. The penultimate nucleotide of the antisense siRNA, however (complementary to position 2 of the 23 nt motif), should always be complementary to the targeted sequence. For simplifying chemical synthesis, TT is always used.

(c) **Pattern filtering**. A series of filters are applied to the SiRNA candidates. Most of these filters are parameters with the recommended values set by default. The values can be changed to other less stringent ones if required.

1. **GC content**. It is strongly recommended to maintain percentages between 26 and 56%, equivalent to the 30–52% suggested by Reynolds et al. (2004), taking into account the leader nucleotides flanking the oligo.
2. **CT content**, with default values between 50 and 90%.
3. **Avoid more than 4 (default) G in a row**. Poly (G) stretches can form aggregates that may interfere with the siRNA binding.
4. **Avoid more than 4 (default) A in a row**.
5. **Avoid more than 4 (default) of the same nucleotide in a row**.
6. **Variation of up to 20% (default) in nucleotide composition**. In this way, the nucleotides do not display extreme percentage values.
7. **Avoid A in position 3**. The pattern NNAN20 is avoided.
8. **After nucleotide 2 (default) set to 3 (default value in the range 2–4) the minimum number of G/Cs in the next 4 bases**.
9. **After nucleotide 17 (default) set to 3 (default value in the range 2–4) the minimum number of A/Ts in the next 4 bases**.
10. **Avoid SNPs**. This option may be useful if the gene is suspected of being polymorphic. In this instance, silencing could be affected by mismatches due to allele variants. This option and the two that follow are active by default.
11. **Avoid exon boundaries**. This option might be useful if alternative splicing can take place in the gene. Avoiding exon boundaries thus reduces the possibility of mispairing siRNA.
12. **Avoid repeat regions**. Again, if repeated regions do exist it is good practice to exclude them.
13. **Discard siRNA candidates with a score below 6 (default)** according to Reynolds et al. (2004) rules. One of these rules regards G+C content, 6 as referring to nucleotides at particular positions and another evaluates the absence of internal repeats. To check for this SiDE uses internally, the program Mold for RNA and DNA secondary structure prediction using nearest neighbour thermodynamic rules (Mathews et al., 1999). The results are used to evaluate the propensity a siRNA has to form internal hairpins.
14. **Discard siRNA candidates with a classification score lower than 1 lb (default) based on Ui-Tei et al. (2004) schema, which is constructed according to the presence of sequence features, such as A/T at particular positions, etc.**
15. **Discard siRNA candidates with less than 2 points (default) of Amarzguioui and Prydz (2004) score, in which the presence of specific features add points to the final score**. An empirical study allowed the authors to define features consistently correlated positively with functionality that include mainly an asymmetry in the stability of the duplex ends (measured as the A/T differential of the three terminal base pairs at either end of the duplex) and the motifs S1, A6 and W19.
16. **Discard siRNA candidates with less than 2 points (default) based on Hsieh et al. (2004) paper**. The authors of this paper do not specifically propose an algorithm to get a score for a particular siRNA, but the sequence features they identify as important for efficient silencing can be used to construct an algorithm. Here, we follow the implementation suggested by Saetrom and Snoe (2004) with an additional 1 for A/T in position 11 as suggested by the authors.
17. **Discard siRNA candidates with less than 10 points (default) based on Takasaki et al. (2004) paper**. The authors developed a selection method based on the gene degradation measure (priority score) defined by positional features of individual nucleotides.

(d) **BLAST filtering**

1. Allow unspecific BLAST alignments with more than 4 (default in the range 1–10) non-identical bases (gap). To include putative siRNA targets in the result if the second unspecific match (if any) has more than 4 gaps.
2. **BLAST alignments specific of Transcript**. Ensembl facilitates the use of transcripts instead of the complete gene (in cases in which different transcripts are available), and siRNA can be selected for silencing specific transcripts but not others.

(e) **Output**

1. **Sorting**. Results can be sorted using different criteria including the melting temperature ($T_m$), differential number of gaps or the different scores according to the distinct methods implemented (Reynolds et al., 2004; Ui-Tei et al., 2004; Amarzguioui and Prydz, 2004; Hsieh et al., 2004; Takasaki et al., 2004). The criteria of a number of mismatches are usually intuitive enough to establish a threshold. Nevertheless, in some situations, such
Fig. 1. Web interface to SiDE with the selection of parameters available.
as genes with biases or strong non-uniformities in base composition, mismatches can result in erroneous estimations of the likelihood of cross-hybridizations. In this instance, the application of criteria that is more related to physical likelihood of matching, such as the melting temperature rather than the simple number of mismatches can be more useful. The $T_m$ of the hybrid siRNA and the target sequence is estimated by using the program MELTING (Le Novère, 2001). Standard values of a sodium concentration of 20 mM and a DNA concentration of 10 μM are used in the calculations. The $T_m$ differential between the specific sequence and the next unspecific candidates can thus be used as a more rational criteria to set a threshold, when avoiding cross-hybridization is the main concern. On the other hand, the use of the different scores is probably more related to functionality of the siRNA.

(2) Remove the first 2 nt at 5′ on Sense and last 2 nt at 3′ on Antisense. This option and the two next ones can be used for convenient representation of the final siRNA sequences.

(3) Remove the first 2 nt at 5′ on Sense and Antisense.

(4) Remove the last 2 nt at 3′ on Sense and Antisense.

Specificity

As previously commented, specificity is probably the most important factor for the successful selection of proper siRNA candidates. If the siRNA match another gene the possibility of silencing side effects cannot be discarded, and conclusions on the functional assay would be impossible.

The idea of specificity comes from the goal of having a unique match in a target gene and reducing the possibility of any possible cross-hybridization in any other gene to a minimum. Specificity is usually achieved by means of using any similarity search strategy, such as BLAST (Altschul et al., 1990), although other alternatives can also be used. Even more important than the algorithm used for the similarity search is the database of reference where the query is raised. There are several technical problems associated with the interpretation of specificity in an automatic way. If the database is redundant and has more than one copy of each gene, a proper parsing of the results is virtually impossible. Having two or more perfect matches is not an indication of possible cross-hybridization. It could be a result of different matches to different versions of the same sequence (or different clones or ESTs). The use of non-redundant databases also poses problems. EST databases do not provide complete information about the genes (typical EST sequences are 200–400 nt long). Databases such as Locuslink contain representative sequences but not always complete genes and some predicted but non-curated genes may be missing. In addition, one gene may have more than one UniGene cluster (identifier) due to the process of clustering ESTs. Therefore, the use of the UniGene database can lead to the rejection of a good siRNA sequence because it corresponds to more than one cluster. This problem also occurs with LocusLink. Other non-redundant databases are automatically compiled with similar criteria or mix genes from different species, making the parsing of the results arduous.

We built the databases for BLAST using the genes annotated in the latest NCBI build genome assembly (in the current implementation, builds 34 for human and 32 for mouse). Sequences are obtained from Ensembl (Birney et al., 2004) and formatted (using formatdb tool) for BLAST use. Each gene is represented by as many transcripts as the annotation in Ensembl describes. Consequently, our databases (one for human and one for mouse genes) contain a single entry for each gene that can be composed by one or more transcript sequences. This allows the selection of gene-specific or transcript-specific siRNAs. Transcript-specific silencing can be useful for the study of processes dependent on alternative splicing. With each new release of Ensembl the databases are rebuilt.

The use of real non-redundant databases allows simple and efficient automatic parsing of the BLAST results and, consequently, makes possible automatic search for siRNA candidates, without requiring human intervention, which is a must in the case of high-throughput experiments.

RESULTS

The procedure for siRNA selection has been implemented as a web-based application (http://side.bioinfo.cnio.es), which is currently available on our website. The use of the program is straightforward, allowing an easy modification of the parameters described above. Figure 2 shows the output of the program. The first line gives information on the version of the database used and allows downloading of the results in the Excel format. The name of the gene scanned is then displayed. If more than one transcript is annotated in Ensembl for a given gene, the resulting siRNA candidates are listed for each transcript. The two first columns refer to the position of the siRNA target in relation to the transcription start point of the transcript. The third column shows the pattern used. There are then two columns with the Sense and Antisense sequence of the siRNA target. The next two columns list the GC% and the melting temperature estimated by means of the program MELTING (Le Novère, 2001). The next five columns list the scores obtained by the different methods mentioned above in the Methods section (Reynolds et al., 2004; Ui-Tei et al., 2002; Amarzguioui and Prydz, 2004; Hsieh et al., 2004; Takasaki et al., 2004), which are related to the likelihood of a siRNA candidate being functional. Arranging siRNA candidates by score is interesting when functionality is the main concern. If avoiding cross-hybridization is the goal, then, the two last columns are the most interesting ones, because they help to select between siRNA candidates when galore. These columns show properties of the next BLAST hit found. When no other BLAST hits are found a dash appears in both columns. This is the most desirable of situations, although more frequently than not, other unspecific hits occur. In this case, the number of gaps with respect to the real hit and the differential of $T_m$ observed can help in the selection of those with a lower likelihood of producing cross-hybridizations. In some cases a melting temperature for some of the next, non-specific BLAST hits cannot be calculated by the program MELTING. This happens because some mismatches produce putative base pairs for which empirical values have not been determined (Mathews et al., 1999) and, in this case, a dash appears in the last column.

If the option BLAST alignments specific of Transcript is checked, no siRNAs are found using the default set of options for this gene, because all of them were common.

Figure 2 shows the candidate siRNAs obtained for the gene APAF1. This gene is involved in the pathways leading to apoptosis. In particular, after caspase activation by cytotoxic stress such as DNA damage, signalling pathways converge on Bcl2 family of proteins, which are involved in the permeabilization of the mitochondria. Then cytochrome $c$ is released and, in a complex with APAF1, activates caspase-9 which continues with the apoptotic process. Among the siRNA candidates found, the one starting at position 1555 (AA T-GTGTCACCTTTACGTGATT) was the siRNA used by several authors to silence the expression of the gene APAF1 (Lassus et al., 2002; Nguyen and Wells, 2003). It is interesting to note that the scores obtained for this siRNA are not particularly high: only 5 for Reynolds et al. (2004), 0 for Amarzguioui and Prydz (2004), 0 for Hsieh et al. (2004), –0.73 for Takasaki et al. (2004) and a class II is obtained using the Ui-Tei et al. (2004) algorithm. Additionally, the likelihood of cross-silencing is negligible: a gap of 10 nt separates this oligo from the next best hit found by BLAST.
High-throughput design of siRNAs

Fig. 2. Output of the program. Candidate siRNAs obtained using the default rules, except the score (Reynolds et al., 2004) which is set to 5. Among the siRNA candidates found, the one starting at position 1555 (the last in the bottom of the figure, corresponding to the oligo AA TTGGTGCACTTTTACGTGA TT) was the siRNA used by several authors to silence the expression of the gene APAF1 (Lassus et al., 2002; Nguyen and Wells, 2003), which in complex with cytochrome c activates caspase-9 and triggers apoptosis.

DISCUSSION

Two important objectives have been addressed in this work that makes SiDE a tool with new features compared with other similar programs: high specificity and flexible use of parameters for optimized prediction of functional siRNA candidates. Both capabilities have been implemented in a way that makes SiDE suitable for high-throughput functional assays using gene silencing. Recently, several siRNA design tools (see Introduction section) have been developed, although none of them help the user to screen for gene-specificity, which constitutes probably the most crucial (and laborious) experimental design steps, in a completely automatic way. Usually, the user must carry out the task of blasting each siRNA candidate sequence individually and then manually examining the resulting (usually several hundreds) sequence alignments. In the best case, specificity is sought by screening for the specified number of mismatches allowed and the length of the match between siRNA and hits, and clustering redundant homologous sequences by grouping hits by LocusLink id or UniGene cluster. As mentioned previously, there are technical problems associated with the interpretation of specificity automatically in these databases. Multiple matches due to different versions of the genes make it virtually impossible to avoid a step of human curation. This makes the systematic use of these programs in

1381
high-throughput projects impossible. In the case of SiDE, the program does not use an already available database. The database has been generated from the gene annotations. In this way (setting aside possible errors of the annotators, which can be considered negligible at the level of known genes), the database is truly non-redundant having a unique copy of each gene. Since automatic processing is then possible, the program can be used for high-throughput siRNA design because it does not require human intervention in the specificity step, which is probably the bottleneck in the design process.

In addition to specificity and energy requirements, there are a number of factors, most of them unknown, that make an siRNA efficient in the silencing process. Recently, different groups undertook systematic analyses of siRNAs [180 siRNAs targeting 2 genes by Reynolds et al. (2004); 148 siRNAs and 30 genes by Hsieh et al. (2004); 62 siRNAs and 6 genes by Ui-Tei et al. (2000) and 46 siRNAs by Amarzguioui and Prydz (2004)] to identify sequence-specific features likely to contribute to the different steps required for efficient silencing. A recent comparative study (Saetrom and Snove, 2004) shows that three score schemes give high and stable performance in the test dataset used: Reynolds et al. (2004), Amarzguioui and Prydz (2004) and Ui-Tei et al. (2004), all of them implemented in SiDE.

To our knowledge, apart from SiDE, only the Dharmaco siDESIGN Center (not suitable for high-throughput scans), includes rules based on these features [in particular, the Reynolds et al. (2004) rules]. Failing to observe these empirical rules will cause a selection of siRNA candidates with considerably lower likelihood of being functional. Given both features together, reliable rules of design and high specificity, SiDE constitutes the most flexible and powerful tool available on the Web for finding siRNA candidates, that can even be used in large-scale gene-silencing functional assays. SiDE can be freely used through its web interface (http://side.bioinfo.cnio.es).

ACKNOWLEDGEMENTS

Special thanks to Amanda Wren for the revision of the English. This work is partly supported by the INB (National Institute of Bioinformatics), funded by Fundación Genoma España. J.M.V. is supported by a FPU fellowship from the MEC.

REFERENCES


