Nucleic Acids Research, 2005, Vol. 33, Database issue ***D131–D134***
doi:10.1093/nar/gki136

siRNAdb: a database of siRNA sequences

Alistair M. Chalk*, Richard E. Warfinge, Patrick Georgii-Hemming and Erik L. L. Sonnhammer

Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius väg 35, S-171 77 Stockholm, Sweden

Received August 15, 2004; Revised and Accepted October 28, 2004

ABSTRACT

Short interfering RNAs (siRNAs) are a popular method for gene-knockdown, acting by degrading the target mRNA. Before performing experiments it is invaluable to locate and evaluate previous knockdown experiments for the gene of interest. The siRNA database provides a gene-centric view of siRNA experimental data, including siRNAs of known efficacy and siRNAs predicted to be of high efficacy by a combination of methods. Linked to these sequences is information such as siRNA thermodynamic properties and the potential for sequence-specific off-target effects. The database enables the user to evaluate an siRNA's potential for inhibition and non-specific effects. The database is available at http://siRNA.cgb.ki.se.

INTRODUCTION

Short interfering RNAs (siRNAs) enable the inhibition of single genes at the nucleotide level. They are duplexes of two RNA molecules, typically 21mers with a 2 nt 3′ overhang (1). A particular strength of this method of knockdown is that an siRNA can be designed to inhibit the expression of any mRNA, and thus the protein it encodes. The knockdown approach, unlike a knockout, allows detailed study of the effects of reducing a gene's expression to none for a period of time, and then allowing its expression to return to normal. This effect can be demonstrated without affecting related proteins, making it an invaluable tool for functional genomics. siRNAs have been found to be effective in *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans* and mammals (2). For an in-depth review of the subject see e.g. (2,3).

With the increased utilization of siRNAs it is essential to keep track of siRNAs that have been published. Experimentalists wish to easily be able to find siRNAs that have already been verified for their target gene. If such siRNAs exist, the researcher will be interested in the conditions under which the siRNA was tested, and the reference to the article where the siRNA was published for further investigation. If no such siRNAs exist then the researcher probably wants to choose an siRNA designed using one of a number of recently published methods (4–8). In both cases it is important to identify potential sequence-specific off-target effects of the siRNA. An additional user group for an siRNA database consists of bioinformaticians. In this case the primary interest is the underlying data, which can be downloaded for subsequent analysis and the building of predictive models.

THE DATABASE

Contents

The database contains information about siRNA molecules from two sources: (i) siRNAs collected from the literature that have experimentally verified efficacy and (ii) siRNAs selected computationally to target the REFSEQ (9) curated human gene set (20 410 'NM' sequences and 6767 'XM' sequences). The database holds experimental information gathered from literature for siRNAs in set (i). This includes efficacy, cell type, efficacy assay and information about the target gene. When exact figures for efficacy are unavailable we approximate the value; these values are marked with a type (validated, predicted, approximated or generalized) to indicate the method used to determine the efficacy value. Detailed descriptions of these types are available online.

For genes with no experimentally verified siRNA in the database, we provide a set of predictions using the following combination of prediction methods. siRNAs were selected only if the siSearch (6) score exceeded 5, the Reynolds (10) score exceeded 5, and the Ui-tei (11) score was Ia or Ib. This set of predicted siRNAs was then subjected to a BLAST specificity search against the REFSEQ database, and siRNAs were retained only if they had no matches to other genes (16 or more consecutive bases).

A link to PubMed with a pre-formulated siRNA query search is also made available to allow the user to easily

*To whom correspondence should be addressed. Tel: +46 7 39824296; Fax: +46 8 337983; Email: alistair.chalk@cgb.ki.se

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

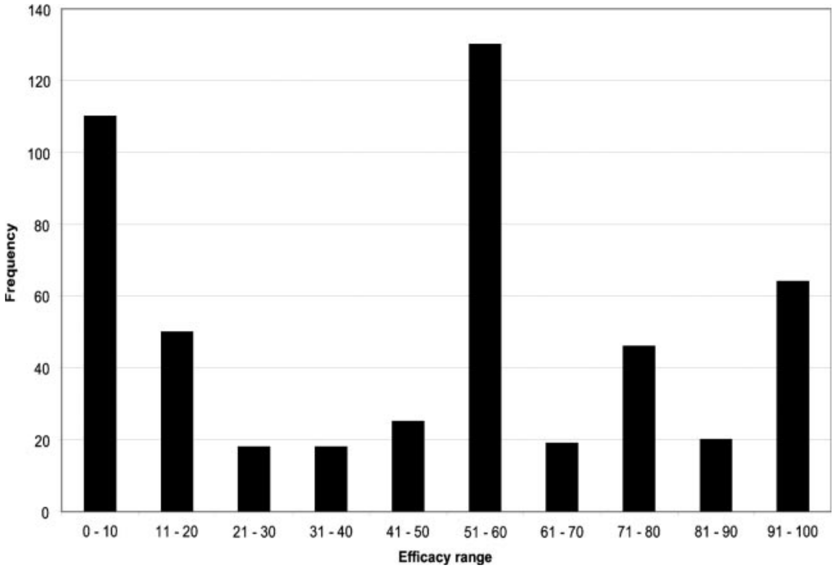


Figure 1. Efficacy level distribution in siRNAdb.

check for new siRNA articles relating to the gene of interest literature.

In release 1.0 there are 500 experimentally verified siRNAs targeting 115 genes. These data were gathered from 55 articles. Since siRNAs are being continuously added to the database in an ongoing manner we recommend checking the server for the latest release information. The distribution of siRNAs per gene is not flat; some genes have a large number of data points, while many contain only a few. The distribution of the efficacies of the siRNAs is shown in Figure 1. Of the 500 siRNAs, 12.8% give knockdown efficacy >90% while 55.8% give efficacy >50%. The experiments use either nucleotide or protein expression levels to measure efficacy; 297 use nucleotide and 198 use protein levels. A number of different transfection reagents were used, with lipofectamine 2000 being the most commonly used (70% of cases) reagent. A total of 109 001 siRNAs (from 21 075 genes) matched the prediction criteria specified in the previous section. Of those, 42 155 siRNAs from 12 888 genes were also found to be specific using BLAST criteria; 14 189 genes have no siRNAs matching these criteria, owing to the strict requirements for the automated predictions. If siRNAdb lacks predictions for a gene it is recommended that the user manually search for siRNA sequences using one, or a combination of siRNA prediction servers.

The database interface

The database interface was designed with the experimental user group as the primary target audience. The user interface is gene-centric, allowing the user to search by nucleotide accession number, free text, sequence or by viewing the list of genes with verified siRNAs.

The layout of the database is straightforward and requires only brief description. For each gene a summary of the siRNAs are shown with links to more detailed information (see Figure 2). The following list illustrates the organization of the data as viewed within siRNAdb. Click on the links below to open the relevant help pages with more verbose documentation than is required here.

- Searching the database
- A gene record
- Individual siRNA records
- Data specific to verified siRNA
- Data specific to predicted siRNA
- Comparison of search methods (AOsearch, BLAST)
- Definitions
- Submit siRNA data to our database
- Downloading the database

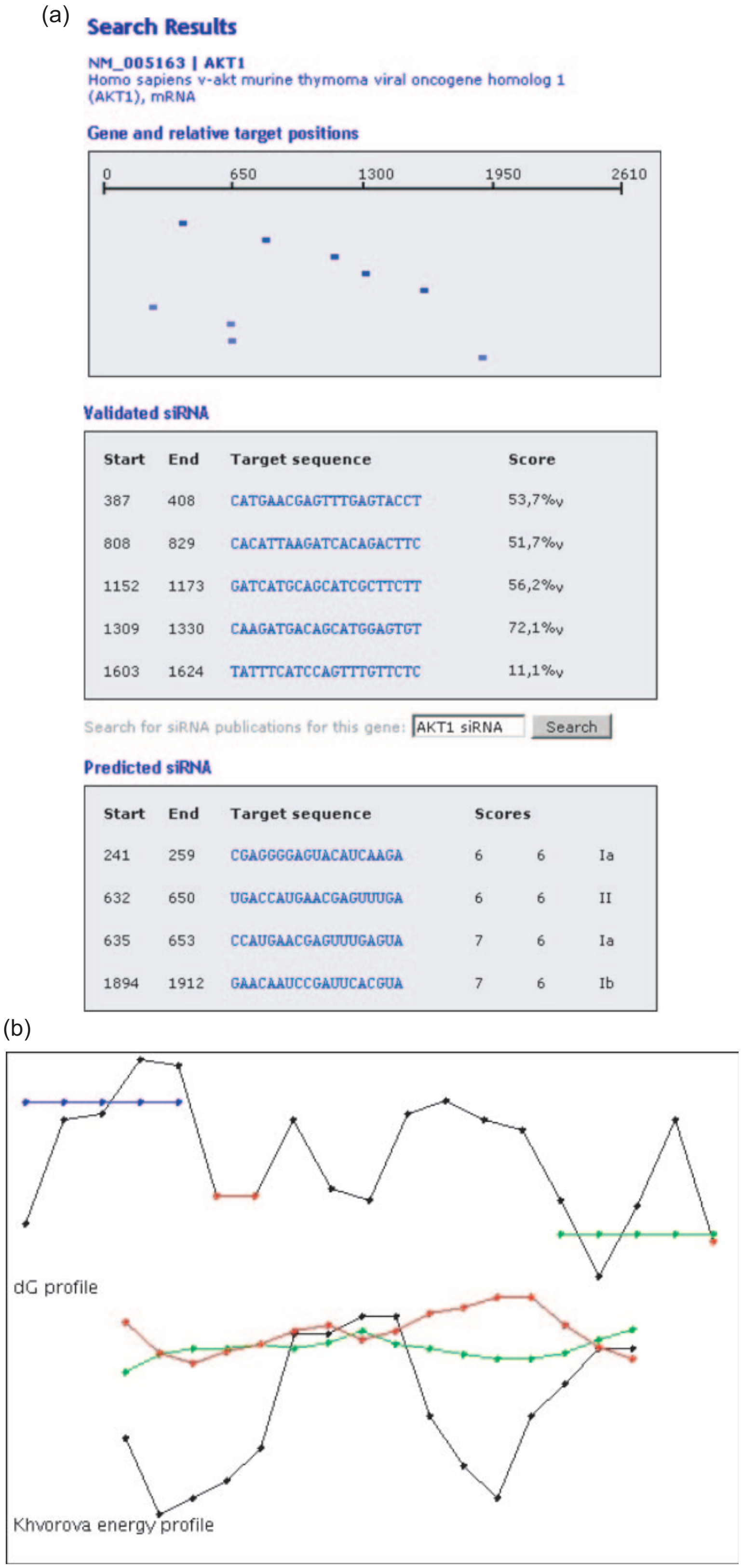
siRNA calculations

A multitude of factors have been identified as being important for governing siRNA efficacy. We calculate and display these factors. Summary statistics are self-evident, and energy profiles are described in (6). The energy data displayed includes the 'start' and 'end' energies, representing the strength of binding at each end of the siRNA.

Implementation

The database is implemented in MySQL version 4. The central table in the database is called *siRNA*, and contains information about the siRNA such as sense and antisense sequences, overhangs and target sequence. The second most important table in the database is the *Experiment* table that contains a list of all experiments performed on the siRNAs as well as references to PubMed. Efficacies for an experiment are stored either as 'validated' or 'predicted' to distinguish these types. Each siRNA can have multiple experiments attached to them, as several experiments can be performed using the same siRNA sequence.

The querying of the underlying SQL database is implemented using Java servlets running on an Apache Tomcat server.



Search Results

NM_005163 | AKT1
Homo sapiens v-akt murine thymoma viral oncogene homolog 1
(AKT1), mRNA

Gene and relative target positions

Validated siRNA

Start	End	Target sequence	Score
387	408	CATGAACGAGTTTGAGTACCT	53,7%$_v$
808	829	CACATTAAGATCACAGACTTC	51,7%$_v$
1152	1173	GATCATGCAGCATCGCTTCTT	56,2%$_v$
1309	1330	CAAGATGACAGCATGGAGTGT	72,1%$_v$
1603	1624	TATTTCATCCAGTTTGTTCTC	11,1%$_v$

Search for siRNA publications for this gene: AKT1 siRNA Search

Predicted siRNA

Start	End	Target sequence	Scores		
241	259	CGAGGGGAGUACAUCAAGA	6	6	Ia
632	650	UGACCAUGAACGAGUUUGA	6	6	II
635	653	CCAUGAACGAGUUUGAGUA	7	6	Ia
1894	1912	GAACAAUCCGAUUCACGUA	7	6	Ib

Figure 2. View of Human AKT1 siRNA data. (**a**) Gene view and (**b**) energy profiles for a single siRNA. The first curve shows di-nucleotide binding energy values, as calculated using the method of Mathews *et al*. (14). The straight horizontal lines represent the binding energy at each end as calculated by Schwarz *et al*. (15). The second set of curves represents free energy profiles calculated using the method of Khvorova *et al*. (16). The black curve is that of the current siRNA. The additional green and red curves are the averaged reference values for best and worst siRNAs, respectively. All curves are calculated from antisense $5' \rightarrow 3'$, which is right → left in this display.

Sequence-specific off target effects

Non-specific off-target effects caused by siRNAs matching genes other than their intended target gene render experimental results hard to interpret or useless. It is essential that siRNAs are designed correctly to take this problem into account. We use two methods for calculating potential sequence-specific off-target effects. For experimentally verified siRNAs we search using AOsearch (http://aosearch.cgb.ki.se) to look for hits with 0–2 mismatches, combined with BLAST. AOsearch uses inexact pattern matching with AGREP (12) that is more sensitive than BLAST for short sequence searches with mismatches. For predicted siRNAs searching with AOsearch is too computationally expensive, hence we used BLAST to identify matches with 16 bp in common. We hope to have AOsearch results incorporated into the server soon, however. For a comparison of exact matching methods versus BLAST (13).

Quality control

We only collect data from other sources and we do not attempt to evaluate the entries ourselves. It is assumed that the experimental verification claimed by the authors of siRNA experiments is correct, and that suitable controls were used to ensure this. By displaying limited experimental information and providing a link to the source article, we provide the user with resources to evaluate the quality of the siRNA.

Submission

In order to maintain an up-to-date resource we encourage experimentalists to submit their siRNA data directly to us as soon as their paper is published. We accept user submissions of siRNA sequences or publications via email. Such submissions are manually checked before addition to the database.

CONCLUSIONS/FUTURE PERSPECTIVES

The database is a collection of siRNA experiments. It was designed to assist experimentalists in determining which siRNA to use to inhibit their gene of interest. As more siRNAs are verified this database will become increasingly useful for developing siRNA design tools. One future plan is to complete a genome-wide siRNA set for the mouse; where human–mouse orthologs are identical, the same siRNA may be used to target both genes. The database was designed to hold results from a number of different prediction methods, and we invite siRNA prediction groups to submit their predictions to the database.

AVAILABILITY

The database implementation and verified siRNAs are available at http://siRNA.cgb.ki.se for non-commercial use. The experimentally verified section of the database is available for download. For-profit organizations are requested to contact the corresponding author.

ACKNOWLEDGEMENTS

We would like to thank Michael McManus and Olga Matveeva for access to their siRNA datasets. This work was supported by a grant from Pfizer Inc., and A.C. was partially funded by a Sweden–Japan foundation stipend.

REFERENCES

1. Elbashir,S.M., Lendeckel,W. and Tuschl,T. (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev*., **15**, 188–200.
2. McManus,M.T. and Sharp,P.A. (2002) Gene silencing in mammals by small interfering RNAs. *Nature Rev. Genet*., **3**, 737–747.
3. Hammond,S.M., Caudy,A.A. and Hannon,G.J. (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nature Rev. Genet*., **2**, 110–119.
4. Naito,Y., Yamada,T., Ui-Tei,K., Morishita,S. and Saigo,K. (2004) siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. *Nucleic Acids Res*., **32**, W124–W129.
5. Yuan,B., Latek,R., Hossbach,M., Tuschl,T., Lewitter,F., Naito,Y., Yamada,T., Ui-Tei,K., Morishita,S., Saigo,K. *et al*. (2004) siRNA Selection Server: an automated siRNA oligonucleotide prediction server. siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. DEQOR: a web-based tool for the design and quality control of siRNAs. *Nucleic Acids Res*., **32**, W130–W134.
6. Chalk,A.M., Wahlestedt,C. and Sonnhammer,E.L. (2004) Improved and automated prediction of effective siRNA. *Biochem. Biophys. Res. Commun*., **319**, 264–274.
7. Henschel,A., Buchholz,F. and Habermann,B. (2004) DEQOR: a web-based tool for the design and quality control of siRNAs. *Nucleic Acids Res*., **32** (Web Server issue), W113–W120.
8. Amarzguioui,M. and Prydz,H. (2004) An algorithm for selection of functional siRNA sequences. *Biochem. Biophys. Res. Commun*., **316**, 1050–1058.
9. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res*., **31**, 34–37.
10. Reynolds,A., Leake,D., Boese,Q., Scaringe,S., Marshall,W.S. and Khvorova,A. (2004) Rational siRNA design for RNA interference. *Nat. Biotechnol*., **22**, 326–330.
11. 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.
12. Wu,S. and Manber,U. (1992) Agrep—a fast approximate pattern-matching tool. *Usenix Winter 1992 Technical Conference*, San Francisco, CA, pp. 153–162.
13. Snove,O.,Jr and Holen,T. (2004) Many commonly used siRNAs risk off-target activity. *Biochem. Biophys. Res. Commun*., **319**, 256–363.
14. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*., **288**, 911–940.
15. Schwarz,D.S., Hutvagner,G., Du,T., Xu,Z., Aronin,N. and Zamore,P.D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.
16. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.