D146–D149 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 7 November 2008*
doi:10.1093/nar/gkn817

siRecords: a database of mammalian RNAi experiments and efficacies

Yongliang Ren, Wuming Gong, Haiyan Zhou, Yejun Wang, Feifei Xiao and Tongbin Li*

Department of Neuroscience, University of Minnesota, Minneapolis, MN 55455, USA

Received September 4, 2008; Revised October 10, 2008; Accepted October 13, 2008

ABSTRACT

RNAi-based gene-silencing techniques offer a fast and cost-effective way of knocking down genes' functions in an easily regulated manner. Exciting progress has been made in recent years in the application of these techniques in basic biomedical research and therapeutic development. However, it remains a difficult task to design effective siRNA experiments with high efficacy and specificity. We present siRecords, an extensive database of mammalian RNAi experiments with consistent efficacy ratings. This database serves two purposes. First, it provides a large and diverse dataset of siRNA experiments. This dataset faithfully represents the general, diverse RNAi experimental practice, and allows more reliable siRNA design tools to be developed with the overfitting problem well curbed. Second, the database helps experimental RNAi researchers directly by providing them with the efficacy and other information about the siRNAs experiments designed and conducted previously against the genes of their interest. The current release of siRecords contains the records of 17 192 RNAi experiments targeting 5086 genes.

INTRODUCTION

RNA interference (or RNAi) is a recently discovered, naturally occurring mechanism for sequence-specific, post-transcriptional down-regulation of gene expression (1). Because RNAi-based gene knockdown techniques (using siRNAs, or small interfering RNAs) offer a fast and cost-effective way of disrupting genes' functions in an easily regulated manner, rapid progress has been made in recent years in the application of these techniques in basic biomedical research and clinical development. In the basic research domain, siRNAs have become a standard gene knockdown tool routinely used in molecular genetics and function genomics laboratories (2,3). In the clinical domain, several RNAi-based therapies against ocular diseases (e.g. AMD or age-related macular degeneration), virus infection (by Hepatitis B and C, and HIV), cancers (e.g. solid tumors) and inflammatory diseases have reached the clinical or pre-clinical trial stage in development (4–6), and a large number of other RNAi-based potential therapeutic agents are actively being explored (7,8).

The successful employment of an RNAi-based gene knockdown technique depends on the proper design or selection of the siRNAs, and the adoption of an effective strategy to deliver the siRNAs to the target cells or tissues (4,9). The purpose of designing siRNAs is to choose from a large number of candidate siRNA sites the ones likely to achieve high potency/efficacy and good specificity (against off-target activity). A properly devised delivery system (using, e.g. viral or non-viral vectors, conjugates, cationic liposomes, or complexes with peptides, polymers, antibodies and aptamers) helps to improve the stability of the siRNA agent, and reduce or eliminate the innate immune response and/or other harmful side-effects induced by the siRNA agent (5,7,10).

The issue of how to design siRNAs that produce high efficacy is the focus of a large body of recent research work [see recent reviews, e.g. (11–16)]. Since it was discovered that not all siRNAs are equally potent in their ability to silence the gene products (17), a series of studies have pointed to a large number of 'features' that might be correlated to the higher efficacy of RNAi experiments. These features can be roughly classified into three categories. The first category are sequence features, including direct sequence features which are defined based on the nucleotide identity in particular positions of the siRNA, e.g. the 6th nucleotide of the siRNA sequence is a 'A' (18,19), and sequence-derived features, e.g. the G/C content of the siRNA is between 30% and 52% (20), and there are no occurrences of more than three identical nucleotides in consecutive positions (21,22). The second category include features defined based on the thermodynamics of the siRNA, e.g. the binding energy in the n7-n11 region is between –1.97 and –1.65 kcal/mol (23), and features surrounding the concept of siRNA duplex terminal

*To whom correspondence should be addressed. Tel: +1 612 3481; Fax: +1 612 626 5009; Email: toli@biocompute.umn.edu

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

asymmetry, e.g. the difference in binding energy between the n16–n19 region and n1–n4 region is greater than 1 kcal/mol (24). The third category of features are defined based on the target sites on the mRNA, including target location-related features, e.g. the target site is outside of the third quartile of the coding region of the mRNA (25), and features focusing on the target site accessibility (26,27), e.g. the local free energy of the most stable structure is greater than or equal to –20.9 kcal/mol (28). Moreover, recent studies suggested that factors related to experimental settings, e.g. the types of siRNA constructs (29,30), the types of cells used (30–34) as well as the methods applied in examining gene products (35) might also influence the efficacy of the RNAi experiments.

A number of siRNA design tools were established in which various combinations of these features were implemented [see recent reviews, e.g. (15,36)]. However, the controversy continues as for which of these features are truly helpful in selecting high-efficacy siRNAs. Meanwhile, it has been increasingly recognized that many earlier siRNA design studies suffered from the 'overfitting' problem (14,37,38)—a term commonly used in the machine learning field, referring to situations where, consequent upon excessive training of a classifier, the performance of the classifier becomes increasingly better on the training data, but worsens on testing data. The only practical way to overcome the overfitting problem is to make use of a large and diverse training dataset (which approximates the ultimate 'testing data'—the general siRNA experimental practice as a whole) when investigating features or factors associated with the higher siRNA efficacy.

We present siRecords (http://siRecords.umn.edu/siRecords), an extensive database of mammalian RNAi experiments with consistent efficacy ratings. Because siRecords hosts the records of all kinds of siRNA experiments conducted with various laboratory techniques and experimental settings, it is a faithful representation of the general, diverse siRNA experimental practice. Recently, using a dataset compiled from siRecords, we analyzed a large number of reported features for their ability to improve RNAi effectiveness. Through carefully combining the most significant features, we derived a bundle of siRNA design rule sets (called the DRM rule sets) which were subsequently shown to outperform a number of established siRNA design tools in selecting effective siRNAs (14). This work demonstrated the usefulness of the siRecords database.

In this article, we outline the design considerations of the siRecords database, its structure and features, and describe the recent improvements made in the siRecords project.

DATABASE CONTENT

siRecords is designed to serve two different purposes: (i) it provides a large and diverse dataset of experimentally validated siRNAs with consistent efficacy ratings, and this dataset can be used by bioinformatics scientists in developing more reliable siRNA design tools, and (ii) it helps experimental RNAi researchers directly by providing the information about what siRNAs have been tested by other researchers against the genes of their interest, and what efficacy levels were achieved in those previous RNAi experiments.

The literature curation and data recording procedures have remained unchanged over the past four years. First, queries are sent to the PubMed database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) for publications related to 'RNAi' and 'siRNA'. Then, the abstracts of the publications are screened, and the full text articles likely to contain information about RNAi gene silencing experiments are retrieved and further examined. Next, for each article containing descriptions of RNAi experiments, the siRNA sequences, the target genes and other key information about experimental conditions are recorded. This information includes: the cells or tissues in which the RNAi experiments were conducted, the forms of the siRNA agents—chemically synthesized oligos or vector transfected shRNAs, and the methods applied in testing the efficacy of the siRNAs—western blot, RT–PCR or others. The siRNA sequences are aligned with the mRNA sequences of the target genes using bl2seq (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi), and the aligned sequences are recorded.

Moreover, an efficacy rating is assigned to each RNAi experiment, based on the description about the result of the gene silencing experiment made in the article. The efficacy rating scheme was designed with balanced considerations. A very coarse-grained rating scheme (for example, a binary scheme that rates siRNAs with 'effective' and 'ineffective') would result in poor usefulness of the database because of the limited information it provides. On the other hand, a very fine-grained rating scheme (for example, one that classifies siRNAs into 10 efficacy categories) would lead to difficulty in obtaining accurate ratings, resulting in a less reliable database being produced. We balanced these two factors and chose to use a four-level rating scheme, where the efficacy of an RNAi experiment is rated as 'very high' if the gene product is reduced by more than 90%; it is rated as 'high' if the gene product is reduced by 70–90%; 'medium' if between 50% and 70% of gene knockdown is achieved; and 'low' if less than 50% of gene knockdown is obtained. The informative sentences in the original articles describing the siRNA efficacy are copied down and stored in the 'original_assessment' field in the database. When adequate textual descriptions about the siRNA efficacy are not available, best efforts are made to assign the efficacy rating scores based on the figures (gel images or summary bar-graphs) presented in the articles, and this information (the basis of the efficacy score assignment) is also kept in the 'original_assessment' field in the database.

During the data deposition process, the siRNA sequence that is maintained in the database may undergo some transformations from the original publication into the database. First, it is possible that DNA bases from the published resource are deposited as RNA, one or more bases represented as 'T' may be transformed into 'U'. Second, it is possible that the sense strand or passenger strand of the siRNA sequence is deposited rather than the guide strand. These are known issues that are being

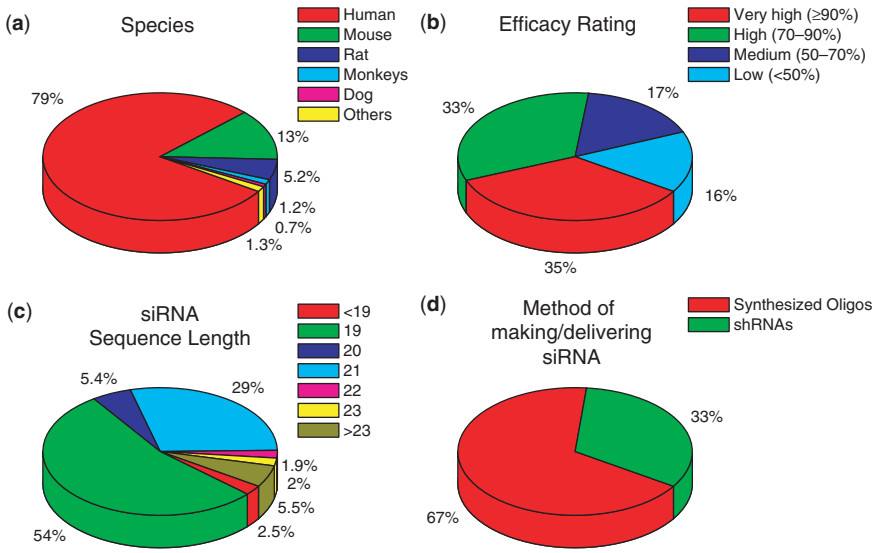

Figure 1. Statistics summary of the records of RNAi experiments in the current release of siRecords. (**a**) Species distribution. The category 'monkeys' includes multiple species, including *Aotus trivirgatus, Cercopithecus aethiops* and *Macca mulatta*. (**b**) Efficacy rating distribution. (**c**) Distribution of the siRNA lengths. (**d**) Distribution of the methods by which the siRNAs are produced and delivered.

actively corrected, but the data are currently heterogeneous as to whether these transformations have occurred or have been corrected. Future releases of siRecords will contain estimates of the degree to which we believe the contents are clean or contain specific kinds of contaminating or transformed data.

There are four major tables in the database schema: SiRecord, which stores the siRNA sequence, key experimental conditions (cell or tissue type, host species, method of making/delivery siRNAs, method of testing efficacy and the test object), original efficacy assessment (sentences related to efficacy assessment in the original articles), and the efficacy rating assigned by siRecords curator; Gene, which stores information about the genes targeted by the siRNAs, including Genbank accession, organism and description of the gene; Correspondent, which stores the contact information of the siRNA origin; and Publication, which stores key information, including PubMed ID and citation data of the original publication.

The current release of siRecords hosts the records of 17 192 RNAi experiments targeting 5086 unique genes, curated from 6122 research articles. The size of the database has more than quadrupled when compared to the first release of the database (Figure 1).

The web interface of the database has recently been rewritten. The improved interface includes a 'siRNA Input Wizard' which will guide data contributors to submit their own records of RNAi experiments with ease. Moreover, the primitive siRNA design tool incorporated in the previous release of siRecords has been replaced by siDRM—a recently developed full-featured siRNA design program in which updated DRM rule sets are implemented (39).

UTILITY

siRecords can be accessed at http://siRecords.umn.edu/siRecords/. At the main page, the user could query a gene by entering the Genbank accession number or GI number, and the matching records would be presented to the user. After the user selects a record, the record display page will present with all relevant information about the record, including the siRNA sequence, experimental setting, efficacy rating and the source of the record. The links to all other records targeting the same gene, and all other records obtained from the same source is displayed.

Data contributors could submit their own records of RNAi experiments with the help of the 'siRNA Input Wizard' shown in the left panel of the web site (registration is required).

DATA ACCESS

The siRecords web site is publically accessible through the URL http://siRecords.umn.edu/siRecords. Academic users can obtain a copy of the current release of the dataset by sending an email to siRecords@biocompute.umn.edu.

IMPLEMENTATION

The siRecords database is a relational database implemented with MySQL on a Fedora II Linux system running on an Intel DUO core 2 computer. The front-end web interface is implemented as a PHP project running under Apache 2.0.

ACKNOWLEDGEMENTS

We thank Q. Xu, X. Zheng, and D. Lin for participating in the curation work in an earlier phase of the project, and the Supercomputing Institute, University of Minnesota for providing computing resources.

FUNDING

University of Minnesota Graduate School and Minnesota Medical Foundation (partial); NIH/NCI (1R21CA126209, 4R33CA126209) (to T.L.). Funding for open access charge: NIH/NCI

Conflict of interest statement. None declared.

REFERENCES

1. Fire,A., Xu,S., 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. Gunsalus,K.C. and Piano,F. (2005) RNAi as a tool to study cell biology: building the genome-phenome bridge. *Curr. Opin. Cell Biol.*, **17**, 3–8.
3. Xia,X.G., Zhou,H. and Xu,Z. (2005) Promises and challenges in developing RNAi as a research tool and therapy for neurodegenerative diseases. *Neurodegener. Dis.*, **2**, 220–231.
4. Kim,D.H. and Rossi,J.J. (2007) Strategies for silencing human disease using RNA interference. *Nat. Rev. Genet.*, **8**, 173–184.
5. Hadj-Slimane,R., Lepelletier,Y., Lopez,N., Garbay,C. and Raynaud,F. (2007) Short interfering RNA (siRNA), a novel therapeutic tool acting on angiogenesis. *Biochimie*, **89**, 1234–1244.
6. de Fougerolles,A., Vornlocher,H.P., Maraganore,J. and Lieberman,J. (2007) Interfering with disease: a progress report on siRNA-based therapeutics. *Nat. Rev. Drug Discov.*, **6**, 443–453.
7. Kuhn,R., Streif,S. and Wurst,W. (2007) RNA interference in mice. *Handbook Exp. Pharmacol.*, 149–176.
8. Gaither,A. and Iourgenko,V. (2007) RNA interference technologies and their use in cancer research. *Curr. Opin. Oncol.*, **19**, 50–54.
9. Inoue,A., Sawata,S.Y. and Taira,K. (2006) Molecular design and delivery of siRNA. *J. Drug Target*, **14**, 448–455.
10. Devi,G.R. (2006) siRNA-based approaches in cancer therapy. *Cancer Gene Ther.*, **13**, 819–829.
11. Kurreck,J. (2006) siRNA Efficiency: Structure or Sequence-That Is the Question. *J. Biomed. Biotechnol.*, **2006**, 83757.
12. Mittal,V. (2004) Improving the efficiency of RNA interference in mammals. *Nat. Rev. Genet.*, **5**, 355–365.
13. Peek,A.S. and Behlke,M.A. (2007) Design of active small interfering RNAs. *Curr. Opin. Mol. Ther.*, **9**, 110–118.
14. Gong,W., Ren,Y., Xu,Q., Wang,Y., Lin,D., Zhou,H. and Li,T. (2006) Integrated siRNA design based on surveying of features associated with high RNAi effectiveness. *BMC Bioinform.*, **7**, 516.
15. Patzel,V. (2007) In silico selection of active siRNA. *Drug Discov. Today*, **12**, 139–148.
16. Matveeva,O., Nechipurenko,Y., Rossi,L., Moore,B., Saetrom,P., Ogurtsov,A.Y., Atkins,J.F. and Shabalina,S.A. (2007) Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. *Nucleic Acids Res.*, **35**, e63.
17. Holen,T., Amarzguioui,M., Wiiger,M.T., Babaie,E. and Prydz,H. (2002) Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. *Nucleic Acids Res.*, **30**, 1757–1766.
18. Amarzguioui,M. and Prydz,H. (2004) An algorithm for selection of functional siRNA sequences. *Biochem. Biophys. Res. Commun.*, **316**, 1050–1058.
19. Takasaki,S., Kotani,S. and Konagaya,A. (2004) An effective method for selecting siRNA target sequences in mammalian cells. *Cell Cycle*, **3**, 790–795.
20. 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.
21. Wang,L. and Mu,F.Y. (2004) A Web-based design center for vector-based siRNA and siRNA cassette. *Bioinformatics*, **20**, 1818–1820.
22. Cui,W., Ning,J., Naik,U.P. and Duncan,M.K. (2004) OptiRNAi, an RNAi design tool. *Comput. Methods Programs Biomed.*, **75**, 67–73.
23. Poliseno,L., Evangelista,M., Mercatanti,A., Mariani,L., Citti,L. and Rainaldi,G. (2004) The energy profiling of short interfering RNAs is highly predictive of their activity. *Oligonucleotides*, **14**, 227–232.
24. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
25. 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 phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic Acids Res.*, **32**, 893–901.
26. Tafer,H., Ameres,S.L., Obernosterer,G., Gebeshuber,C.A., Schroeder,R., Martinez,J. and Hofacker,I.L. (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.
27. Shao,Y., Chan,C.Y., Maliyekkel,A., Lawrence,C.E., Roninson,I.B. and Ding,Y. (2007) Effect of target secondary structure on RNAi efficiency. *RNA*, **13**, 1631–1640.
28. Schubert,S., Grunweller,A., Erdmann,V.A. and Kurreck,J. (2005) Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. *J. Mol. Biol.*, **348**, 883–893.
29. McManus,M.T., Haines,B.B., Dillon,C.P., Whitehurst,C.E., van Parijs,L., Chen,J. and Sharp,P.A. (2002) Small interfering RNA-mediated gene silencing in T lymphocytes. *J. Immunol.*, **169**, 5754–5760.
30. McManus,M.T. and Sharp,P.A. (2002) Gene silencing in mammals by small interfering RNAs. *Nat. Rev. Genet.*, **3**, 737–747.
31. Elmaagacli,A.H., Koldehoff,M., Peceny,R., Klein-Hitpass,L., Ottinger,H., Beelen,D.W. and Opalka,B. (2005) WT1 and BCR-ABL specific small interfering RNA have additive effects in the induction of apoptosis in leukemic cells. *Haematologica*, **90**, 326–334.
32. Nicholson,L.J., Philippe,M., Paine,A.J., Mann,D.A. and Dolphin,C.T. (2005) RNA interference mediated in human primary cells via recombinant baculoviral vectors. *Mol. Ther.*, **11**, 638–644.
33. Guan,R., Tapang,P., Leverson,J.D., Albert,D., Giranda,V.L. and Luo,Y. (2005) Small interfering RNA-mediated Polo-like kinase 1 depletion preferentially reduces the survival of p53-defective, oncogenic transformed cells and inhibits tumor growth in animals. *Cancer Res.*, **65**, 2698–2704.
34. Spankuch-Schmitt,B., Bereiter-Hahn,J., Kaufmann,M. and Strebhardt,K. (2002) Effect of RNA silencing of polo-like kinase-1 (PLK1) on apoptosis and spindle formation in human cancer cells. *J. Natl Cancer Inst.*, **94**, 1863–1877.
35. Atkinson,P.J., Young,K.W., Ennion,S.J., Kew,J.N., Nahorski,S.R. and Challiss,R.A. (2006) Altered expression of G(q/11alpha) protein shapes mGlu1 and mGlu5 receptor-mediated single cell inositol 1,4,5-trisphosphate and Ca(2+) signaling. *Mol. Pharmacol.*, **69**, 174–184.
36. Birmingham,A., Anderson,E., Sullivan,K., Reynolds,A., Boese,Q., Leake,D., Karpilow,J. and Khvorova,A. (2007) A protocol for designing siRNAs with high functionality and specificity. *Nat. Protoc.*, **2**, 2068–2078.
37. Chalk,A.M., Wahlestedt,C. and Sonnhammer,E.L. (2004) Improved and automated prediction of effective siRNA. *Biochem. Biophys. Res. Commun.*, **319**, 264–274.
38. Saetrom,P. and Snove,O. (2004) A comparison of siRNA efficacy predictors. *Biochem. Biophys. Res. Commun.*, **321**, 247–253.
39. Gong,W., Ren,Y., Zhou,H., Wang,Y., Kang,S. and Li,T. (2008) siDRM: an effective and generally applicable online siRNA design tool. *Bioinformatics*, **24**, 2405–2406.