D108–D111 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki131

HuSiDa—the human siRNA database: an open-access database for published functional siRNA sequences and technical details of efficient transfer into recipient cells

Matthias Truss*, Maciej Swat[2], Szymon M. Kielbasa[2], Reinhold Schäfer[1], Hanspeter Herzel[2] and Christian Hagemeier

Department of Pediatrics, Laboratory for Molecular Biology and [1]Laboratory of Molecular Tumor Pathology, Institute of Pathology, Charité, Universitätsmedizin–Berlin, Germany and [2]Institute for Theoretical Biology, Humboldt-University, Berlin, Germany

Received August 15, 2004; Revised and Accepted October 25, 2004

ABSTRACT

Small interfering RNAs (siRNAs) have become a standard tool in functional genomics. Once incorporated into the RNA-induced silencing complex (RISC), siRNAs mediate the specific recognition of corresponding target mRNAs and their cleavage. However, only a small fraction of randomly chosen siRNA sequences is able to induce efficient gene silencing. In common laboratory practice, successful RNA interference experiments typically require both, the labour and cost-intensive identification of an active siRNA sequence and the optimization of target cell line-specific procedures for optimal siRNA delivery. To optimize the design and performance of siRNA experiments, we have established the human siRNA database (HuSiDa). The database provides sequences of published functional siRNA molecules targeting human genes and important technical details of the corresponding gene silencing experiments, including the mode of siRNA generation, recipient cell lines, transfection reagents and procedures and direct links to published references (PubMed). The database can be accessed at http://www.human-siRNA-database.net. We used the siRNA sequence information stored in the database for scrutinizing published sequence selection parameters for efficient gene silencing.

INTRODUCTION

During the last three years, a new technology has revolutionized functional genomics. Transfection of small interfering RNA (siRNA) molecules into mammalian cells induces post-transcriptional gene silencing via sequence-specific mRNA degradation (RNA interference) (1,2).

siRNA molecules for RNAi experiments can be generated by chemical or by enzymatic synthesis (1,3). Alternatively, expression vectors can be employed that encode short hairpin RNAs. Expressed hairpin molecules are intracellularly processed into functional siRNA by the ribonuclease Dicer (2–4). Only a fraction of randomly chosen 19 bp siRNA sequences has the capability of inducing efficient gene silencing (5). Therefore, gene knock down experiments typically require a labour and cost-intensive optimization of potent siRNAs and protocols for efficient siRNA delivery into the cell lines of interest that are frequently refractory toward siRNA incorporation. The availability of functional siRNA sequences and siRNA transfection protocols is steadily increasing due to the rapidly growing number of published applications of RNAi. This prompted us to establish the human siRNA database (HuSiDa) (http://www.human-siRNA-database.net). The database serves as a repository for both, sequences of published functional siRNA molecules targeting human genes and important technical details of the corresponding gene silencing experiments. It aims at supporting the setup and actual procedure of specific RNAi experiments in human cells.

THE DATABASE

Data acquisition and database overview

Published articles describing functional siRNA sequences in PubMed are identified by a query with the keywords RNAi, RNA interference and siRNA. The effectiveness of each siRNA sequence was judged by its gene silencing activity as shown in the results section of the publication. siRNA

*To whom correspondence should be addressed. Tel: +49 3 0450 566197; Fax: +49 3 0450 566913; Email: Matthias.Truss@Charite.de sequences were only approved if their silencing effects significantly exceeded 50%. In addition, we verified whether each siRNA matches its target gene mRNA. Mismatches or lack of a corresponding mRNA sequence are automatically detected and indicated in the HuSiDa. Each database record consists of the target gene name including the corresponding gene accession numbers (UniGene, RefSeq), the siRNA sequence (sense strand), its activity in the percentage of gene silencing (if available), manuscript reference (PubMed), the cell line and transfection method used, the siRNA source and the target gene description.

Recently, it has been proven that 15 (perhaps sometimes even as few as 11) contiguous nucleotides are sufficient for directing the degradation. This implies that a given siRNA may elicit effects even on mRNAs that were not intended to be down-regulated but additionally targeted because of sequence similarities (6). To address this issue, we employed the RefSeq database search to identify the length of the longest contiguous part of the siRNA that in addition to the target mRNA also matches other mRNA sequences. This value, becoming larger when chances for off-target effects increase, we term specificity and we calculate it for all siRNAs.

Data retrieval

To retrieve functional siRNAs sequences for a specific gene, the human siRNA database can be searched for UniGene cluster IDs, gene names, RefSeq accession numbers and target gene descriptions via a web interface (www.human-siRNA-database.net) downloads of the complete database content are available. Another important feature of the siRNA database is the availability of technical background information as a basis for the design of RNAi experiments. For instance, the option to search for cell lines enables the user to readily identify suitable siRNA transfection protocols.

Current status and future developments

The database is updated on a weekly basis. Currently, the database encompasses 1158 entries. Owing to the enormous popularity of RNA interference in research and experimental therapy, the availability of new siRNA sequences is growing exponentially. We therefore strongly encourage the submission of siRNA sequence datasets by users, provided that the sequences have been accepted for publication in a peer-reviewed journal. For this purpose, we have generated a web form for data submission.

The number of tolerable matches in siRNA sequences is still under debate. Typically, the popular BLAST algorithm is used for estimating target specificity. Owing to the known limitations of sequence alignment, many published siRNAs may lead to sequence-dependent off-target effects (7). Therefore, in future versions of the database, calculation of the specificity will consider position-dependent target mismatch tolerances.

DATABASE ANALYSIS

Several attempts have been made for identifying sequence characteristics associated with highly effective siRNAs molecules. Statistical analysis of the correlation between siRNA sequence and efficacy has led to the development of several siRNA design algorithms (5,8–14). They are based on various criteria including sequence features, hairpin formation potential, duplex stability and secondary structure features of mRNA. A recently published comparative analysis revealed significant qualitative differences in the performance of these algorithms. Three out of four of the well-performing algorithms are based on sequence features; the best performing one is based on weighted sums of sequence motifs and patterns (15). This suggests that the relevant information required for the efficient prediction of siRNAs efficacy is contained in its primary sequence and that the relevance of flanking sequences and mRNA secondary structure have probably been overrated in the past. It is important to note that all previous experiments for sequence optimization have been restricted to synthetic siRNA molecules. Reports indicating that siRNA sequences sometimes lose silencing activity when embedded in a short hairpin format (16) suggest that the key determinants of silencing activity may be distinct in the short hairpin context.

We have performed a statistical analysis of active siRNA sequences from HuSiDa. Only siRNA sequences of 19 bp length and validated silencing activity have been included to ensure a homogeneous, well defined dataset. The dataset DB_all consists of 603 sequences, targeting 396 different genes; the dataset DB_sh (short hairpin) consists of 87 active short hairpin siRNA sequences (see Supplementary Material). The statistical analysis includes (i) the analysis of position-dependent base preferences within the 19 bp siRNA sequence core and the 5′ and 3′ flanking dinucleotides, (ii) global GC-content, (iii) the exclusion of base repetitions and (iv) the existence of stable hairpin structures.

We first analyzed the frequencies of concurrencies of G or C nucleotides at specific positions of the 19 bp siRNA core of the 603 active siRNA sequences listed in our database (Figure 1). This analysis revealed a strong preference for G or C nucleotides at position 1. This is in agreement with the findings of Ui-Tei *et al.* (10), Reynolds *et al.* (5) and Takasaki *et al.* (8). In addition, positions 4, 7 and 10 preferentially harbor G or C nucleotides, while G or C nucleotides at position 6 were underrepresented. However, the significance of these sequence preferences is questionable. Approximately 80% of the published siRNA sequences have been designed according to the recommendations of Elbashir *et al.* (1). Consequently, these sequences reside in the coding regions of their target genes and follow AA dinucleotides. A statistical analysis of nucleotide preferences of random 19mers and random 19mers that follow AA dinucleotides is displayed in Figure 1 E and F. Selection for preceeding AA dinucleotides precludes sequence preferences for G or C nucleotides at positions 1, 4, 7, 10, 13 and 16. This 3 bp periodicity is due to different base compositions within reading frames (17). For instance, the first position in most reading frames has an excess of G nucleotides. If the rule of preceding AA dinucleotides is applied, the resulting 19 mers are preferably in the reading frame and, consequently, G nucleotides are overrepresented at positions 1, 4, 7, etc. This could, at least partially, contribute to the observed preferences for G or C nucleotides at positions 1, 4, 7 and 10 in the selected siRNA population (Figure 1D).

Similar to these observations are the results published by Takasaki *et al.* (8). However, in our opinion, these results are due to the siRNA design rules, as described above, and are not an underlying principle of functional siRNAs.

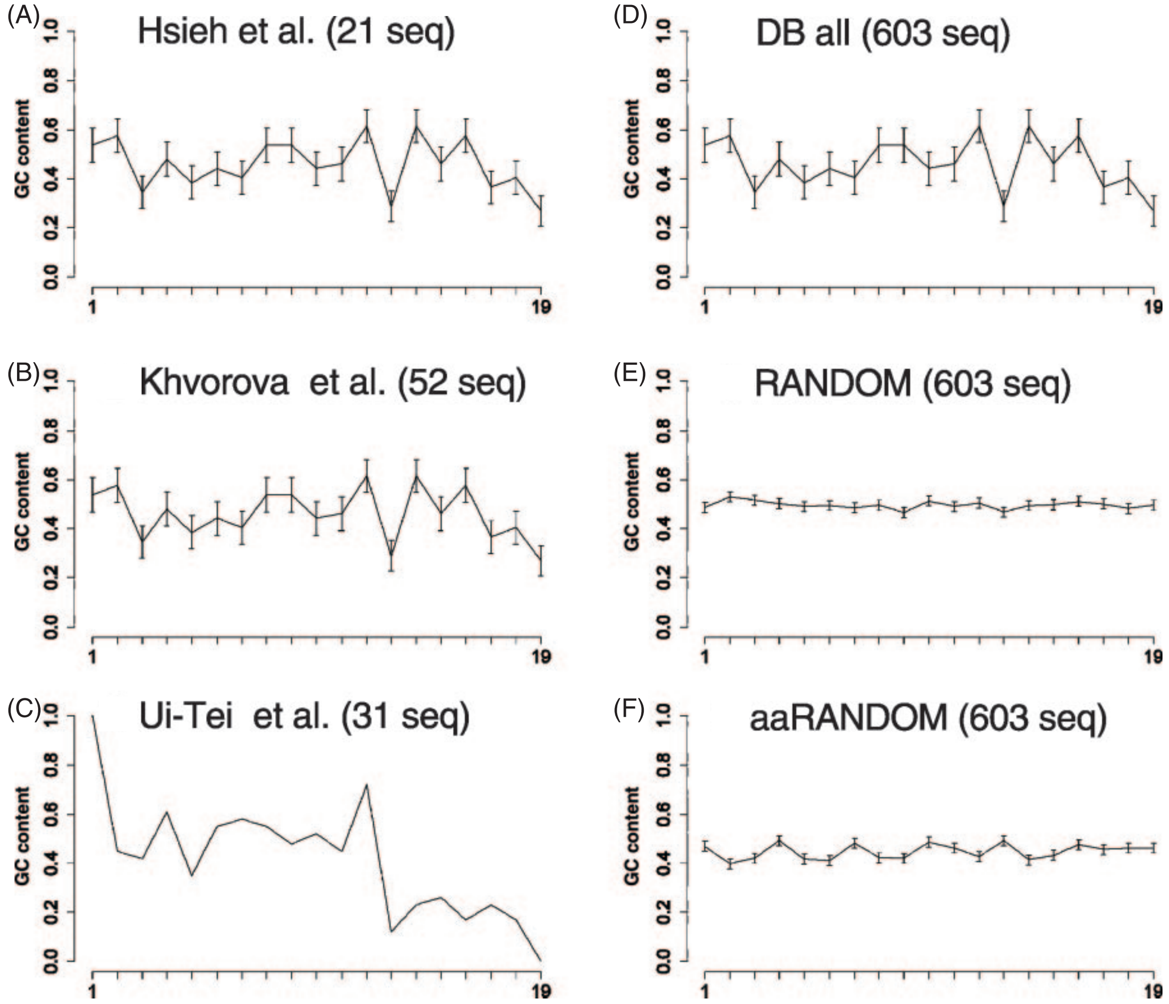

Figure 1. Comparison of frequencies of G or C base occurrences of active or randomly selected siRNA sequences. Each position of the 19mer siRNA core starting at the 5′ end of the sense strand was analyzed. Sequences were obtained from: (**A**) Hsieh *et al*. (9), (**B**) Reynolds *et al*. (5), (**C**) Ui-Tei *et al*. (10) and (**D**) from our the HuSiDa listing 603 active siRNA sequences. Controls: (**E**) 603 randomly selected 19mers and (**F**) 603 19mers that follow an AA dinucleotide.

The analysis of sequence position-dependent preferences for G or C nucleotides in DB_all and the subset of short hairpin siRNA sequences (DB_sh) gave comparable results (see Supplementary Material).

Reynolds *et al*. (5) have identified eight criteria (I–VIII) associated with functional siRNA molecules. For example, to match criterium II, siRNA sequences must exhibit a low internal stability of the 3′ end of the sense strand. This is the case if the sense strand harbors at least 3 ‘A/U’ bases at positions 15–19. A detailed statistical analysis of our DB_all dataset (603 sequences) revealed that sequences matching this criterion are not over but slightly underrepresented (47.1% of active versus 52.2%), when compared to a set of randomly selected 19mers (see Supplementary Materials). The same holds true for most of the other criteria. It would be interesting to analyze subsets of highly active siRNAs (>80% silencing) instead of the complete database (>50% silencing). However, at present, the number of sequence entries with assigned well-defined silencing efficacies is still limiting.

It is our intention to facilitate the refinement of siRNA design algorithms by providing a collection of rapidly accessible siRNA sequences. Consequently, downloads of the complete HuSiDa dataset are available.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

M.S. and S.M.K. are supported by the German Federal Ministry of Education and Research (BMBF) and the German Research Foundation (DFG). M.T., R.S. and C.H. are supported by a grant (NGFN KR-S01T06) from the German Federal Ministry for Research and Education (BMBF).

REFERENCES

1. Elbashir,S.M., Harborth,J., Lendeckel,W., Yalcin,A., Weber,K. and Tuschl,T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.

2. Yu,J.Y., DeRuiter,S.L. and Turner,D.L. (2002) RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells. *Proc. Natl Acad. Sci. USA*, **99**, 6047–6052.
3. Paddison,P.J., Caudy,A.A., Bernstein,E., Hannon,G.J. and Conklin,D.S. (2002) Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev*., **16**, 948–958.
4. Brummelkamp,T.R., Bernards,R. and Agami,R. (2002) A system for stable expression of short interfering RNAs in mammalian cells. *Science*, **296**, 550–553.
5. 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.
6. Jackson,A.L., Bartz,S.R., Schelter,J., Kobayashi,S.V., Burchard,J., Mao,M., Li,B., Cavet,G. and Linsley,P.S. (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol*., **21**, 635–637.
7. Snove,O.,Jr and Holen,T. (2004) Many commonly used siRNAs risk off-target activity. *Biochem. Biophys. Res. Commun*., **319**, 256–263.
8. Takasaki,S., Kotani,S. and Konagaya,A. (2004) An effective method for selecting siRNA target sequences in mammalian cells. *Cell Cycle*, **3**, 790–795.
9. 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.
10. 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.
11. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
12. Chalk,A.M., Wahlestedt,C. and Sonnhammer,E.L. (2004) Improved and automated prediction of effective siRNA. *Biochem. Biophys. Res. Commun*., **319**, 264–274.
13. Luo,K.Q. and Chang,D.C. (2004) The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochem. Biophys. Res. Commun*., **318**, 303–310.
14. Saetrom,P. (2004) Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinformatics*, in press.
15. Saetrom,P. and Snove,O.,Jr (2004) A comparison of siRNA efficacy predictors. *Biochem. Biophys. Res. Commun*., **321**, 247–253.
16. Karpilow,J., Leake,D. and Marshall,B. (2004) siRNA: enhanced functionality through rational design and chemical modification. *PharmaGenomics*, 32–40.
17. Holste,D., Grosse,I., Buldyrev,S.V., Stanley,H.E. and Herzel,H. (2000) Optimization of coding potentials using positional dependence of nucleotide frequencies. *J. Theor. Biol*., **206**, 525–537.