Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 319 (2004) 264–274
BBRC
www.elsevier.com/locate/ybbrc

Improved and automated prediction of effective siRNA

Alistair M. Chalk,* Claes Wahlestedt, and Erik L.L. Sonnhammer

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

Received 26 April 2004

Abstract

Short interfering RNAs are used in functional genomics studies to knockdown a single gene in a reversible manner. The results of siRNA experiments are highly dependent on the choice of siRNA sequence. In order to evaluate siRNA design rules, we collected a database of 398 siRNAs of known efficacy from 92 genes. We used this database to evaluate previously proposed rules from smaller datasets, and to find a new set of rules that are optimal for the entire database. We also trained a regression tree with full cross-validation. It was however difficult to obtain the same precision as methods previously tested on small datasets from one or two genes. We show that those methods are overfitting as they work poorly on independent validation datasets from multiple genes. Our new design rules can predict siRNAs with efficacy $\geqslant$50% in 91% of cases, and with efficacy $\geqslant$90% in 52% of cases, which is more than a twofold improvement over random selection. Software for designing siRNAs is available online via a web server at http://sisearch.cgb.ki.se/ or as a standalone version for high-throughput applications.

RNAi is a recently discovered biological phenomenon whereby a single gene may be inhibited at the RNA stage of synthesis. short interfering RNAs (siRNAs) are duplexes of two RNA molecules, typically 21-mers with a 2nt 3′ overhang [1]. One strand is loaded into the RISC complex [2] after which a sequence specific cleavage of the target takes place. The strength of the principle behind any form of RNA targeting (siRNA and antisense) is that the molecule can be used to inhibit expression of any mRNA, and thus the protein it encodes. 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 [3]. Many reviews of the biological processes behind siRNA inhibition exist, see for example [3,4].

Efficient reliable design of high efficacy siRNA molecules is essential to meet the needs for cost-effective high-throughput functional genomics projects. To meet these needs the siRNAs designed should at least conform to the following criteria: (a) be predicted with high accuracy, (b) be sequence specific, and (c) be produced in a form that facilitates high-throughput production. In this paper we address these criteria, focusing primarily on a and b.

Despite the apparent ease of designing siRNAs (compared to a popular DNA-based knockdown technique: antisense oligonucleotides (AOs)), a number of problems still remain. Randomly selected siRNAs produce knockdown $\geqslant$50% with 58–78% success rate, while very effective siRNAs ($\geqslant$90/95%) are found by chance 11–18% of the time [5,6].

Initial design paradigms for siRNAs were based on motif rules, such as AAN(19)TT, with 3′ TT overhangs on both strands. These evolved into motifs to match siRNAs with strong binding at one end (the sense 5′ end). Recent studies have demonstrated that binding energy is a key factor in siRNA design. The more sophisticated methods compare binding energies along the siRNA molecule. In brief, findings indicate that relative and absolute binding energies of the 5′ antisense and 5′ sense strands determine which strand enters the RISC complex [7]. The energy values for positions 9–15 have also been shown to be important in siRNA design [6]. Other studies have indicated that low GC content is correlated with high efficacy [8,9].

* Corresponding author.
E-mail address: alistair.chalk@cgb.ki.se (A.M. Chalk).

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.04.181

Further studies stress the importance of 5′ terminal nucleotides for determining siRNA efficacy. Antisense 5′ AU-rich, sense 5′ GC-rich, and antisense positions 1–7 AU-rich are criteria introduced by [10]. Point specific nucleotide preferences for positions 6, 11, 13, 16, and 19 have also been indicated [11]. A recent study uses a rule-based approach, incorporating positional rules with criteria such as %GC content and hairpin formation potential [5]. Additional common design guidelines include avoiding the following types of motifs: AAAA/TTTT, GGG/CCC, and long stretches of GCs (e.g., CCGCGGC, GCGCGGCG).

Synthetic synthesis of siRNAs is not cost-effective for large-scale screening projects. This problem has been addressed by using expression vectors as delivery method. These vectors produce stable amounts of siRNA utilizing the cell's own machinery. Mammalian expression vectors synthesizing siRNA-like transcripts

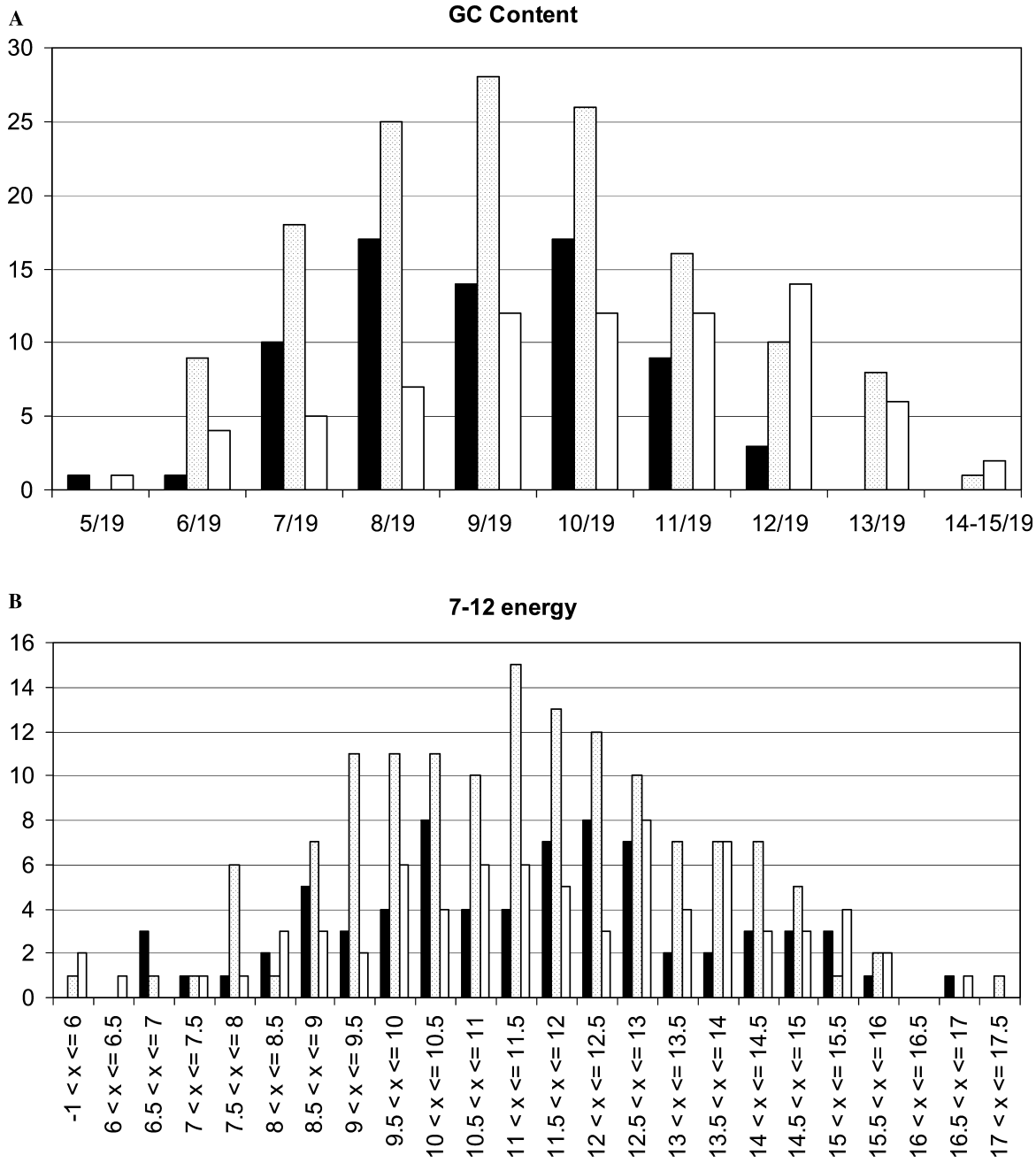


Fig. 1. Distribution of siRNAs in each efficacy category for the given property. The column coloring of black (■) dotted (▨), and white (□) represents siRNAs with efficacy high (≥90%), moderate (≥50%,<90%), and low (<50%), respectively. (A) %GC content. (B) Middle energy (7–12). (C) Total hairpin energy (sense + antisense strands). (D) Difference between antisense and sense 5′ energy. (E) Antisense 5′ energy. (F) Sense 5′ energy. All 5′ energies as calculated as in [7].

are able to cause gene knockdown [12]. Short hairpin (shRNAs) are another example of a similar technique [13].

There are currently a number of computational tools available for siRNA design. The tools almost exclusively look at base composition and position relative to the start codon [14–16]. Most allow user-defined motifs, but do not allow constraints on binding energy. Other features commonly present are screening for non-specific siRNAs using BLAST [17] and the removal of siRNAs spanning SNP sites.

In this paper, we evaluate current design paradigms using a newly assembled siRNA database. We then evaluate the potential parameter space and describe a novel classification method based on rules applied to a subset of these parameters. Using our new method, which is based primarily on thermodynamic properties, we achieved an improved success rate. We present software implementing our method for designing specific high efficacy siRNAs. On the Web server the method is integrated with common alternative design methods and graphical display of the results.

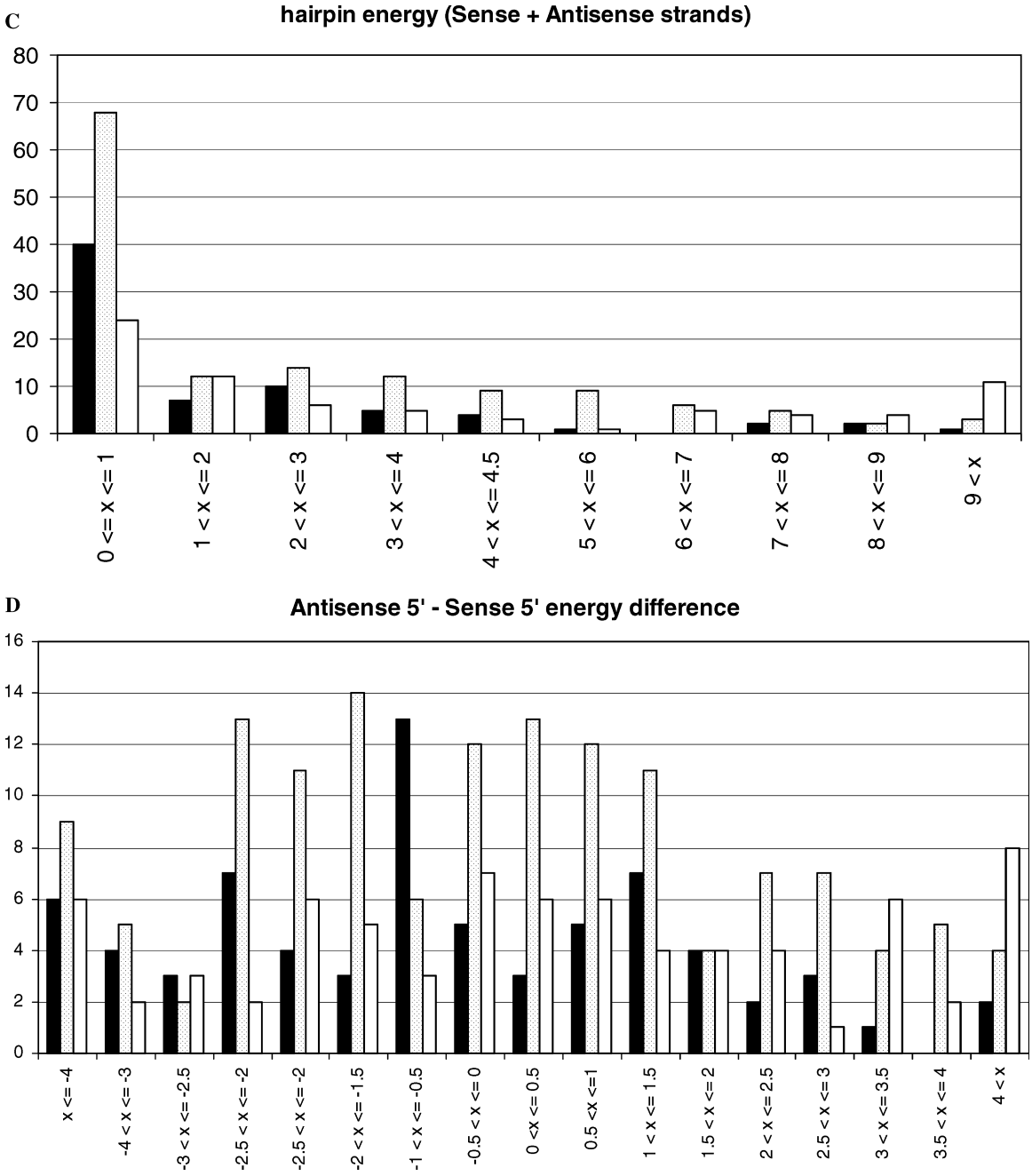


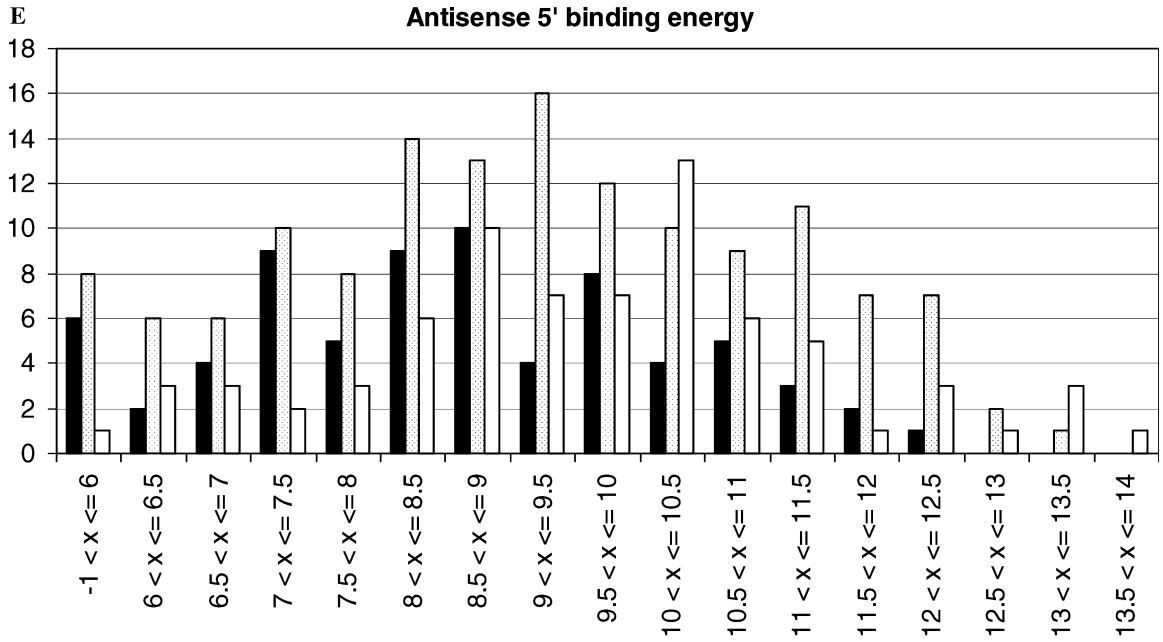
Fig. 1. (*continued*)

Materials and methods

Databases. The database used for this work was derived from a literature search for articles containing functional siRNAs. The database consists of 30 studies and contains 398 siRNAs targeting 92 genes. For the analysis only siRNAs of length 19 (duplex region) were used (287 siRNAs—17 studies, 30 genes) [6,18–33]. Expression levels are either at the RNA or protein level. For this analysis we use the same system as [6] for efficacy classification: high, moderate, and low ($\geqslant 90\%$, $90\% < x \geqslant 50\%$ and $<50\%$, respectively). The 287 siRNAs were distributed in these categories the following way: high: 25%, moderate: 49%, and low: 26%.

Calculations. siRNA nomenclature: base-numbering is referred to by the sense strand (same as target mRNA). The first nucleotide of the 5′ end of the sense strand is position 1; this corresponds to position 19 of the antisense strand. Thermodynamic properties for siRNAs were calculated at 39 °C using free energy tables [34], except for energy profiles from [6] where the method is described. Sense and antisense 5′ end energy comparisons were calculated using methods described in [7]. Hairpin energy is calculated using the Vienna RNA package at 39 °C [35]. All energy values are given as absolute values in units of kcal/mol. GC content is calculated using the 19-mer double-stranded region of the siRNA. T_m calculations for predicting the melting temperature of the hairpin loop were calculated based on the nearest neighbor model [36] using the Oligo 6.0 software package.

Statistical analysis. Statistical tools from Excel (Microsoft) were used for scatter plot data presentations and p value calculations. Re-



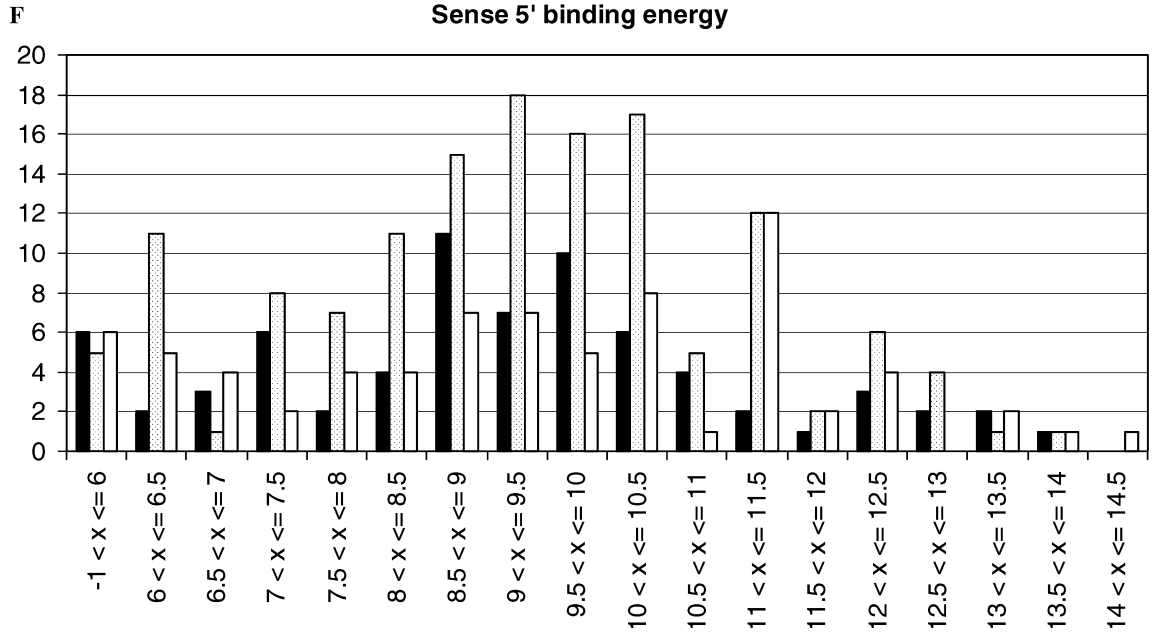


Fig. 1. (*continued*)

gression trees were produced using the R statistical package, using the rpart library (http://www.r-project.org).

Software. Summary data for siRNAs were generated using si-Search, an application written in Java to support siRNA design. The application requires BioJava 1.3.1 (http://www.biojava.org) and Java Runtime Environment 1.4.2. A website interface to this software is available at http://sisearch.cgb.ki.se/ and a command line version of the software is available on request.

Results and discussion

In order to evaluate existing design methods and develop novel methods we compiled a database of siRNAs (see Materials and methods). Evaluation using a database of siRNAs from many sources allows the validation of models on a number of genes from different experimental groups. This allows better generalization, as using a small number of genes may bias the results if those genes are more or less susceptible to certain siRNA features than the average gene. The distribution of efficacy in the database of 287 siRNAs is 25%, 49%, and 26% (high, moderate, and low, respectively).

We used the database to analyze the effect of siRNA efficacy-determining features proposed in the literature. As expected, regions exist for most features where the number of low efficacy siRNAs is overrepresented (Fig. 1). For example, siRNAs with high GC content (53% or above) are low efficacy in 22/44 of cases (Fig. 1A). Also, siRNAs with GC content less than 36% appear enriched in low efficacy cases. High binding energy (>13 kcal/mol) in the middle region of the siRNA (positions 7–12) results in a higher proportion of low efficacy siRNAs (Fig. 1B), and siRNAs with total hairpin energy of one or more are less likely to be effective siRNAs (Fig. 1C).

It has been suggested by Schwarz et al. [7] that one of the most important factors of siRNA efficacy is the binding energy difference between the antisense 5′ and sense 5′ ends. If the binding energy is stronger at the antisense 5′, the incorrect strand is supposed to be loaded into the RISC complex, and no RNA interference will result. However, when we investigated this energy difference in isolation, we found no strong correlation with siRNA efficacy (Fig. 1D), although a trend towards low efficacy is visible for differences above 2 kcal/mol. The antisense 5′ and sense 5′ end binding energies are shown alone in Figs. 1E and F, and here a trend is visible for low energy to correlate with high efficacy. To investigate this further we plotted antisense 5′ energy vs. sense 5′ energy (Fig. 2). The majority of high efficacy siRNAs have a sense 5′ end binding energy stronger than the antisense 5′ end. Low efficacy siRNAs are broadly spread across all energy differences except for extreme values ($\leqslant -4$, >3) where 10/30 siRNAs had low efficacy. In both Figs. 1D and 2 there is a pronounced bias for high efficacy siRNAs in the region −1 and 0 kcal/mol antisense 5′–sense 5′ energy difference (just above the diagonal in Fig. 2). This may be due to some bias in our database, although the 18 high efficacy (out of 44) siRNAs in this region came from five different studies. Most of them came from the two genes studied by Khvorova et al. [6] but these were screening studies with no preselection of target sites. Still, because this rule might not be entirely general we have only included it in our design scheme as an addition to the more general rule that the energy difference should be <0.

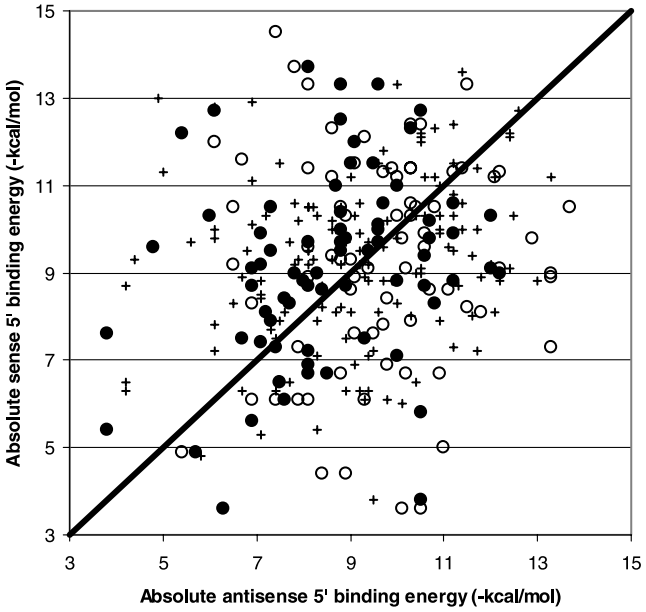


Fig. 2. Antisense vs. sense end binding energy. High, moderate, and low efficacies are represented by solid circles, hollow circles, and plus signs, respectively.

A thorough statistical analysis was performed using a comprehensive set of potential energy parameters.

Table 1
Significant ($p < 0.001$) measures for differentiating between high and low efficacy siRNAs

Parameter	Start base	End base	p value (high vs. low)
Energy*	18	19	<0.00001
Energy	17	19	<0.00001
Energy*	4	8	0.00005
Energy	4	7	0.00005
Energy*	4	6	0.00006
Energy	16	19	0.00009
Energy*	5	8	0.00009
Energy*	6	7	0.00011
Energy	3	8	0.00012
Energy	3	7	0.00021
Energy	1	7	0.00022
S 5′ energy			0.00032
Energy	2	7	0.00052
%GC content			0.00064
Energy	15	19	0.00091

Student's t test was performed comparing the high and low efficacy siRNA sets for each parameter.

* Indicates also significant ($p < 0.001$) for low vs. moderate.

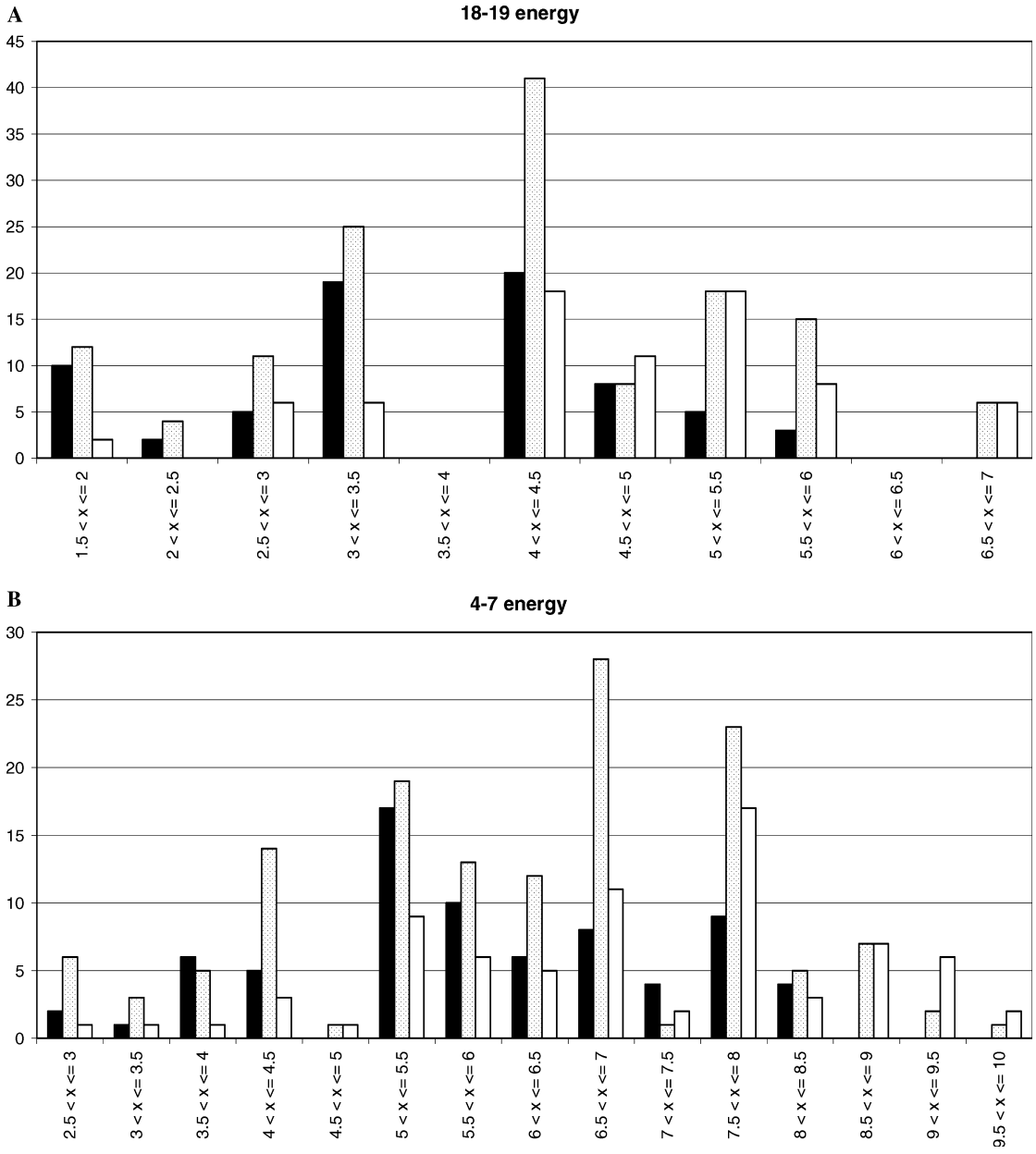


Fig. 3. Distribution of siRNAs in each efficacy category for properties with highest statistical significance. The column coloring of black, dotted, and white represents siRNAs with efficacy high (≥90%), moderate (≥50%, <90%), and low (<50%), respectively. (A) 18–19 energy, (B) 4–7 energy, (C) 4–8 energy, and (D) 4–7 energy using Khvorova et al. [6] method.

We analyzed binding energy parameters covering all 1–6-mer regions of the antisense strand of the siRNA, plus previously discussed parameters and energy parameters described by Khvorova et al. [6]. In total 129 parameters (120 energy parameters, three bond count parameters, three hairpin energy parameters, and three GC content parameters) were tested and the most significant ($p < 0.001$) parameters for high vs. low efficacy siRNAs are shown in Table 1. The regions where the binding energy was statistically significant overlap considerably, falling into two main categories. The first describes positions in the region 15–19, the second describing positions 1–9 (with the strongest signal in the region 4–9). The effects of these significant parameters are shown in Fig. 3. Analysis of low vs. moderate data produced five significant ($p < 0.001$) parameters that overlapped with the 15 significant parameters in the low vs. high analysis. However, when comparing moderate vs. high efficacy siRNAs none of the parameters produced a significant p value (<0.001). Single factors are thus not sufficient to separate moderate from high efficacy siRNAs.

To evaluate existing methods for siRNA design we used the siRNA database. The scoring method of

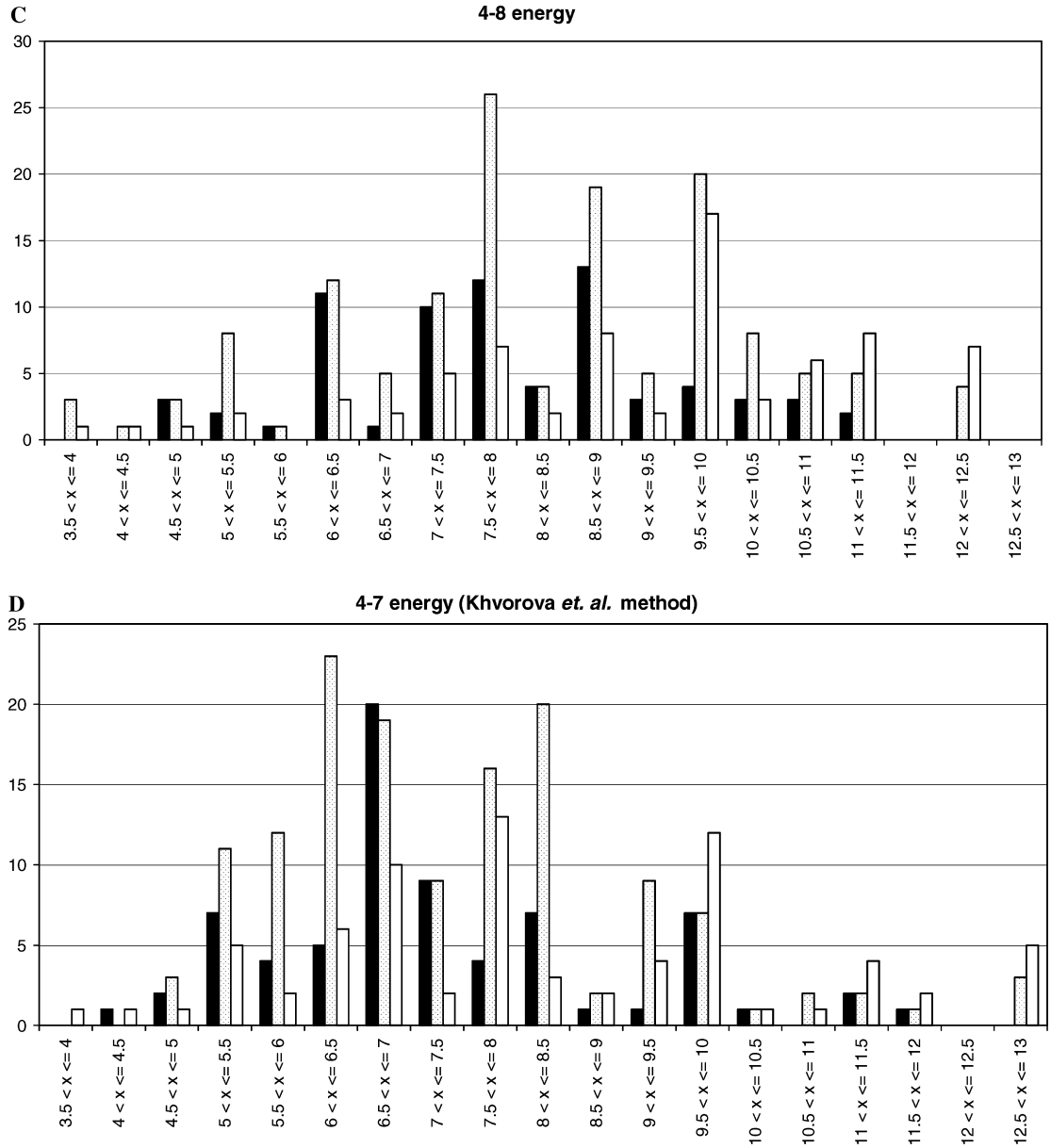


Fig. 3. (*continued*)

Table 2
Current rules and parameters with highly significant p values

Rule	siRNAs conforming	Ratio high/low	p value
18–19 energy	102	2.57	<0.00001
4–7 energy	50	2.00	0.00005
4–8 energy	172	1.77	0.00005
AS 5′ energy	133	1.69	0.74463
Energy difference	58	1.38	0.03096
%GC content	241	1.31	0.00064
Middle energy	189	1.22	0.45277
S 5′ energy	214	1.19	0.00032
Hairpin	199	1.16	0.18639
Khvorova 4–7 energy	220	0.96	0.00195
Whole dataset	N/A	0.41	N/A

Ratio high/low is the number of high efficacy siRNAs divided by the number low efficacy siRNAs remaining after removing non-conforming siRNAs.

Reynolds et al. [5] was tested using a split dataset to remove bias (indicating which siRNAs the method used for training). Their method performs significantly better on their training set (33/65 correctly predicted high) than on an independent test set (5/16), which can be caused by overfitting to the training data. This type of result is typical of strategies using a small number of genes for training, and shows the need for verification of strategies on larger and more varied test sets.

Based on our analysis we created a scoring scheme for evaluating siRNAs. For each of the parameters in Fig. 1 we chose cutoffs that define regions that maximize the ratio of high/low efficacy siRNAs. The derived rules (termed ‘Stockholm rules’) are:

(a) Total hairpin energy < 1.
(b) Antisense 5′ end binding energy < 9.
(c) Sense 5′ end binding energy in range 5–9 exclusive.
(d) GC between 36% and 53%.
(e) Middle (7–12) binding energy < 13.
(f) Energy difference < 0.
(g) Energy difference within -1 and 0.

The score of an siRNA is incremented by one for each rule fulfilled, giving a score range of (0,7). The effect of each parameter on the ratio of high/low efficacy siRNAs is shown in Table 2. The result of applying the Stockholm scoring scheme to the database is shown in Fig. 4. There is a clear relationship between the score and the relative percentages of low, moderate, and high efficacy siRNAs. Using this scoring scheme and a cutoff score of 6, the proportion of high efficacy siRNAs is 52%, which is a twofold enrichment compared to the entire database.

To further explore the parameter space we utilized the regression (classification) tree technique [37]. Regression trees were designed based on the total parameter set to classify siRNAs into the classes low, moderate, and high. To avoid overfitting, we generated the trees using the Khvorova et al. [6] dataset and tested the resulting trees on the remainder of the database. The best results were obtained for trees with low efficacy examples up-weighted by a factor of 2, which had a 44% (48/108) success rate of classifying high efficacy siRNAs (Fig. 5, right panel). This is less than the best rule-based methods, but on the other hand none of them were cross-validated. However, the tree method performs better than the other methods when predicting low efficacy siRNAs (Fig. 5, left panel). Surprisingly, the energy parameters previously found as most statistically significant were not chosen by the tree method. This may be due to the stronger classification power of the base rules (e.g., A or U at position 19), modeling moderate data points, or simply that the 'old' energy rules do not hold up under cross-validation.

An assessment of the performance of different siRNA prediction methods applied to the siRNA database is shown in Fig. 5. The methods can be split into two classes: (a) low-performance methods [10,38], with a success rate of 25–30% on high efficacy siRNAs, and (b) high-performance methods ([5,6], Stockholm rules, Tree) with a success rate of 44–52% on high efficacy siRNAs. The same trend is found when excluding the Khvorova data from the database, although that leaves too few datapoints (106) to interpret reliably. In practice, a user would want to combine the results of different high-performing methods when selecting siRNAs. In order to test this technique one would need more data however.

The specificity of the siRNA is an important aspect of the design process. siRNAs can potentially hybridize with multiple RNA target genes and produce misleading experimental results [39]. To identify potential cross-hybridization we run BLAST against cDNA databases such as Unigene [40] or Ensembl [41] without any filtering. As a guideline, 19 bp siRNA sequences with identical matches of length $\geqslant 16$ to other genes should be discarded. A search for potential miRNAs is also recommended; however, BLAST has limited sensitivity for this type of search and

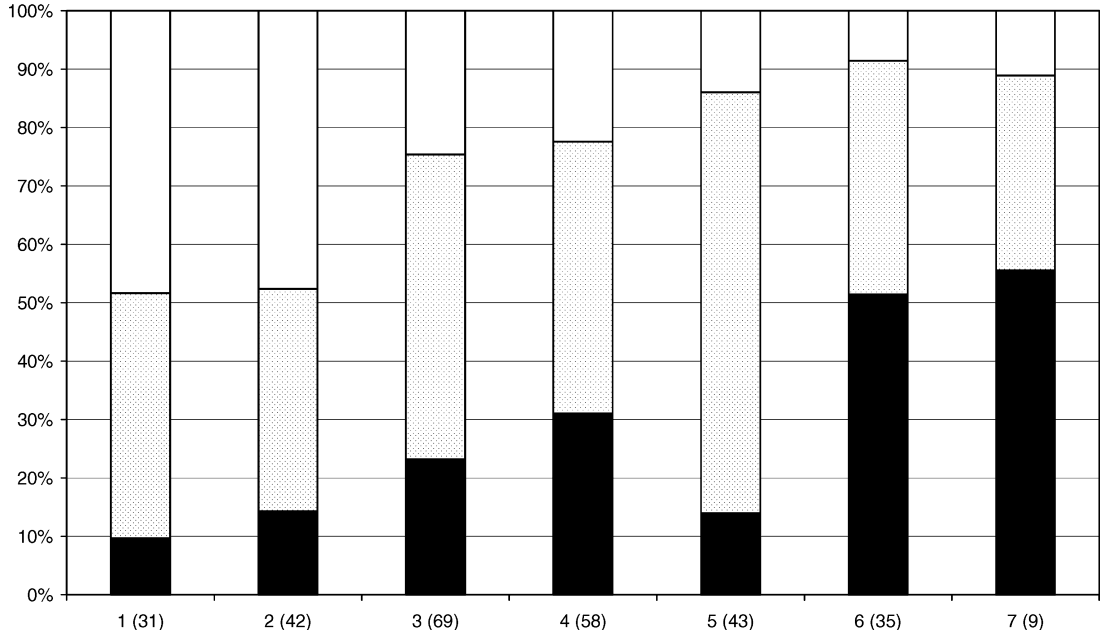


Fig. 4. Classifying the siRNA db using the Stockholm rules. Numbers in brackets indicate number of examples in each score class.

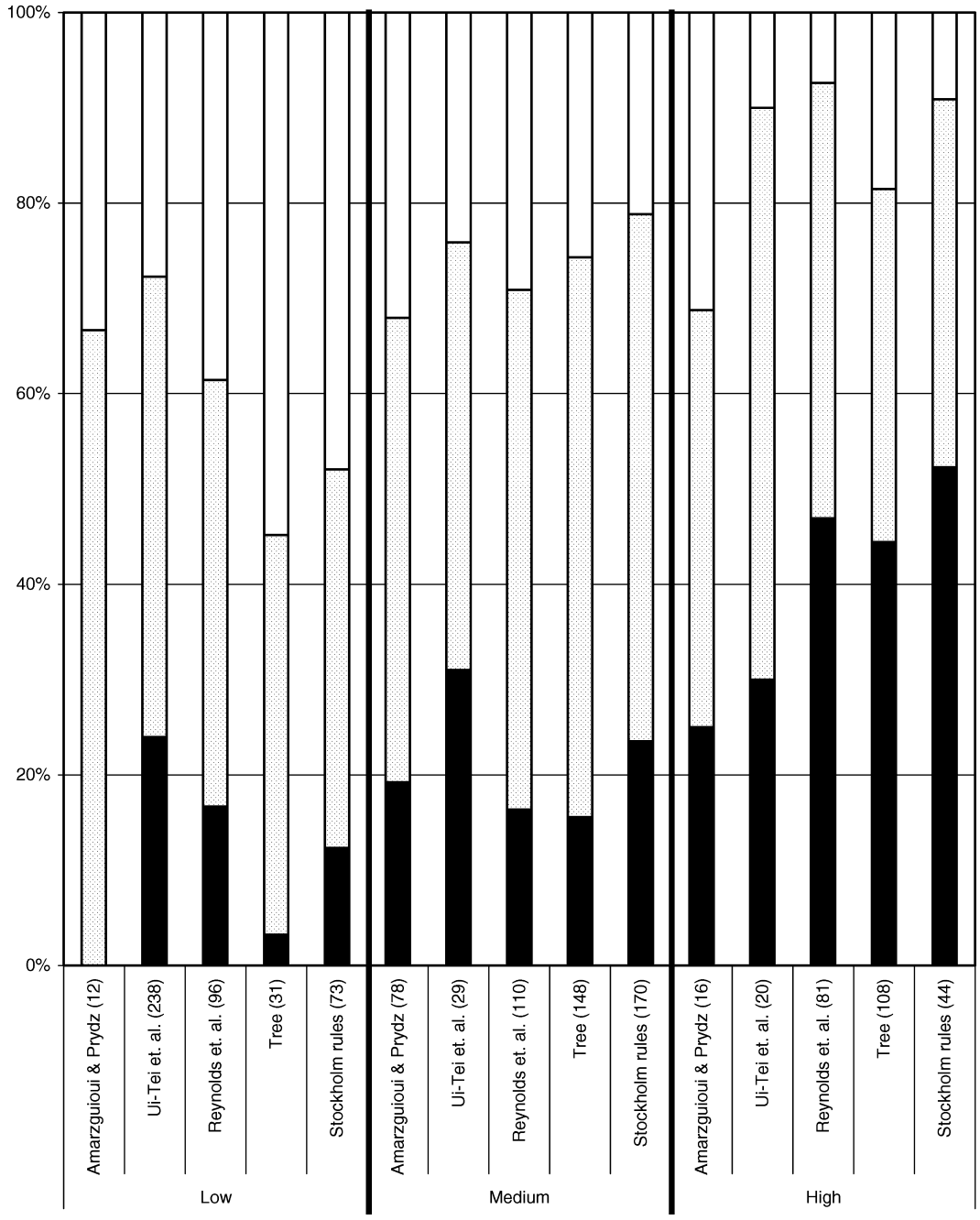


Fig. 5. Performance of methods on siRNA database. Score based methods are split as follows (notation low, moderate, and high): Reynolds et al. (0–2, 3–4, and 5+), Amarzguioui and Prydz (−4 to −2, −1 to 2, and 3 to 5), Ui-tei et al. (II and III, Ib, Ia), and Stockholm rules (0–2, 3–5, and 6–7).

should only be used for a quick analysis where a more exact search is not feasible.

siSearch

The siSearch software tool is a flexible siRNA design program implemented in Java. Through a web or command line interface (http://www.sisearch.cgb.ki.se) the user can select the following combinations of design rules for selecting siRNAs:

1. %GC content of the siRNA.
2. Stockholm rules score.
3. Regression tree classification—trained on the Khvorova et al. dataset.
4. Reynolds et al. rules score (without the oligo 6.0 T_m calculation).
5. Ui-tei et al. rules score.
6. Amarzguioui and Prydz rules score.
7. Special motifs (AAN[19], AAN19TT, NARN[17]YNN, and custom motifs).

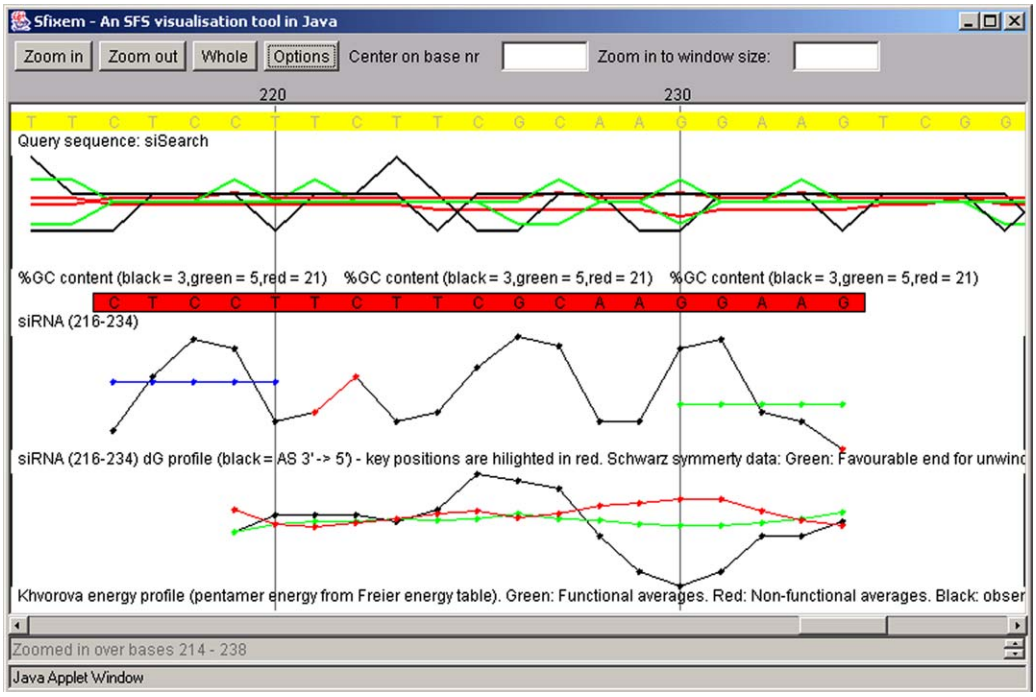


Fig. 6. siSearch output viewed in Sfixem. The curves directly under the siRNA (216–234) segment show di-nucleotide binding energy (dG) values, calculated according to Mathews et al. [34]. The binding energy of each end calculated by the Schwarz et al. method is shown as horizontal lines. Positions 6, 7, and 19 are considered the most important parts of the curve for siRNA design and are shown in red. The lower set of curves are the free energy profiles as calculated by Khvorova et al. [6]. Averaged reference energy curves from their sets of best (green) and worst (red) siRNAs are displayed, which can be compared to the measured curve for this siRNA. At position 230, the measured energy is the lowest curve, while the 'worst siRNA' curve is the highest. All curves are calculated from antisense $5' \rightarrow 3'$, which is right $\rightarrow$ left in the Sfixem display. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

8. Disallow certain motifs (AAAA/TTTT, CCC/GGG, and long stretches of consecutive GC).
9. Schwarz et al. energy difference between sense and antisense ends.

Additional built-in resources include BLAST searching in order to ensure that the siRNA is specific to the target gene. The user may specify the stringency required for a match. It is also possible to siRNAs that are valid both in the target organism and a corresponding model organism. The user provides a homologous sequence to screen against, providing a search for siRNAs valid across multiple species.

Output options include ordering data for the siRNAs and vector models described previously. The output is in html or the more graphically oriented SFS format [42], viewable with Sfixem [43] (see Fig. 6). The graphical representation displays the binding energy profile of the siRNAs, in addition to diagnostic measures used for the siRNA design. Using the SFS format allows additional data from other sources (such as repeat regions, homology results) to be added if desired.

Design of siRNAs is a fast-changing field; design rules will continually change as our knowledge of the field increases. Our software allows the combination of a number of approaches described previously; by using this strategy siRNAs that agree with many known design rules can be found. The system is easily extendable to include new rules as they are discovered.

Conclusions

We have evaluated siRNA design by analyzing individual parameters and known design methods using a database of siRNAs. Following statistical analysis we developed a novel scoring scheme and trained a regression tree classification mechanism, improving the theoretical success rate of rationally designed siRNAs. Using the new design rules would have produced inhibition $\geqslant$50% in 91% of the cases predicted as highly effective, and an inhibition $\geqslant$90% in 52% of the same cases.

We have developed software for selecting siRNAs that is flexible and allows the user to choose between or combine a number of methods for siRNA design. The software is available online for single gene siRNA design and is also available as a standalone application suitable for high-throughput requirements.

Acknowledgments

We would like to acknowledge Joacim Elmén, Håkan Thonberg, and Hong-Yan Zhang for their feedback during development. The siRNA database of Michael McManus was invaluable. This work was supported by a grant from Pfizer Inc.

References

[1] S.M. Elbashir, W. Lendeckel, T. Tuschl, Genes Dev. 15 (2001) 188–200.
[2] J. Martinez, A. Patkaniowska, H. Urlaub, R. Luhrmann, T. Tuschl, Cell 110 (2002) 563–574.
[3] M.T. McManus, P.A. Sharp, Nat. Rev. Genet. 3 (2002) 737–747.
[4] S.M. Hammond, A.A. Caudy, G.J. Hannon, Nat. Rev. Genet. 2 (2001) 110–119.
[5] A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W.S. Marshall, et al., Nat. Biotechnol. 22 (2004) 326–330.
[6] A. Khvorova, A. Reynolds, S.D. Jayasena, Cell 115 (2003) 209–216.
[7] D.S. Schwarz, G. Hutvagner, T. Du, Z. Xu, N. Aronin, P.D. Zamore, Cell 115 (2003) 199–208.
[8] T. Holen, M. Amarzguioui, M.T. Wiiger, E. Babaie, H. Prydz, Nucleic Acids Res. 30 (2002) 1757–1766.
[9] S.M. Elbashir, J. Harborth, K. Weber, T. Tuschl, Methods 26 (2002) 199–213.
[10] K. Ui-Tei, Y. Naito, F. Takahashi, T. Haraguchi, H. Ohki-Hamazaki, A. Juni, R. Ueda, K. Saigo, Nucleic Acids Res. 32 (2004) 936–948.
[11] A.C. Hsieh, R. Bo, J. Manola, F. Vazquez, O. Bare, A. Khvorova, S. Scaringe, W.R. Sellers, Nucleic Acids Res. 32 (2004) 893–901.
[12] T.R. Brummelkamp, R. Bernards, R. Agami, Science 296 (2002) 550–553.
[13] P.J. Paddison, A.A. Caudy, E. Bernstein, G.J. Hannon, D.S. Conklin, Genes Dev. 16 (2002) 948–958.
[14] P. Rice, I. Longden, A. Bleasby, Trends Genet. 16 (2000) 276–277.
[15] Research, W.I.f.B. (2003).
[16] N. Levenkova, Q. Gu, J.J. Rux, Bioinformatics 20 (2004) 430–432.
[17] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. 25 (1997) 3389–3402.
[18] N. Ancellin, C. Colmont, J. Su, Q. Li, N. Mittereder, S.S. Chae, S. Stefansson, G. Liau, T. Hla, J. Biol. Chem. 277 (2002) 6667–6675.
[19] J. Carnes, M. Jacobson, L. Leinwand, M. Yarus, RNA 9 (2003) 648–653.
[20] H.K. Chung, Y.W. Yi, N.C. Jung, D. Kim, J.M. Suh, H. Kim, K.C. Park, J.H. Song, D.W. Kim, E.S. Hwang, et al., J. Biol. Chem. 278 (2003) 28079–28088.
[21] J. Harborth, S.M. Elbashir, K. Bechert, T. Tuschl, K. Weber, J. Cell Sci. 114 (2001) 4557–4565.
[22] J.E. Heinonen, C.I. Smith, B.F. Nore, FEBS Lett. 527 (2002) 274–278.
[23] A.L. Jackson, S.R. Bartz, J. Schelter, S.V. Kobayashi, J. Burchard, M. Mao, B. Li, G. Cavet, P.S. Linsley, Nat. Biotechnol. 21 (2003) 635–637.
[24] T. Katome, T. Obata, R. Matsushima, N. Masuyama, L.C. Cantley, Y. Gotoh, K. Kishi, H. Shiota, Y. Ebina, J. Biol. Chem. 278 (2003) 28312–28323.
[25] M. Kullmann, U. Gopfert, B. Siewe, L. Hengst, Genes Dev. 16 (2002) 3087–3099.
[26] C.D. Novina, M.F. Murray, D.M. Dykxhoorn, P.J. Beresford, J. Riess, S.K. Lee, R.G. Collman, J. Lieberman, P. Shankar, P.A. Sharp, Nat. Med. 8 (2002) 681–686.
[27] Y. Peng, Q. Zhang, H. Nagasawa, R. Okayasu, H.L. Liber, J.S. Bedford, Cancer Res. 62 (2002) 6400–6404.
[28] G. Randall, A. Grakoui, C.M. Rice, Proc. Natl. Acad. Sci. USA 100 (2003) 235–240.
[29] E. Song, S.K. Lee, J. Wang, N. Ince, N. Ouyang, J. Min, J. Chen, P. Shankar, J. Lieberman, Nat. Med. 9 (2003) 347–351.
[30] D.R. Sorensen, M. Leirdal, M. Sioud, J. Mol. Biol. 327 (2003) 761–766.
[31] U.N. Verma, R.M. Surabhi, A. Schmaltieg, C. Becerra, R.B. Gaynor, Clin. Cancer Res. 9 (2003) 1291–1300.
[32] T.A. Vickers, S. Koo, C.F. Bennett, S.T. Crooke, N.M. Dean, B.F. Baker, J. Biol. Chem. 278 (2003) 7108–7118.
[33] S.K. Wong, D.W. Lazinski, Proc. Natl. Acad. Sci. USA 99 (2002) 15118–15123.
[34] D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner, J. Mol. Biol. 288 (1999) 911–940.
[35] I.L. Hofacker, Nucleic Acids Res. 31 (2003) 3429–3431.
[36] D.R. Groebe, O.C. Uhlenbeck, Nucleic Acids Res. 16 (1988) 11725–11735.
[37] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed., Springer, Berlin, 2002.
[38] M. Amarzguioui, H. Prydz, Biochem. Biophys. Res. Commun. 316 (2004) 1050–1058.
[39] F. Vazquez, S.R. Grossman, Y. Takahashi, M.V. Rokas, N. Nakamura, W.R. Sellers, J. Biol. Chem. 276 (2001) 48627–48630.
[40] D.L. Wheeler, D.M. Church, S. Federhen, A.E. Lash, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, E. Sequeira, T.A. Tatusova, et al., Nucleic Acids Res. 31 (2003) 28–33.
[41] M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, et al., Nucleic Acids Res. 31 (2003) 38–42.
[42] E.L. Sonnhammer, J.C. Wootton, Proteins 45 (2001) 262–273.
[43] A.M. Chalk, M. Wennerberg, E.L. Sonnhammer, Bioinformatics 20 (2004) (in press).