nature biotechnology

The impact of target site accessibility on the design of effective siRNAs

Hakim Tafer[1,5], Stefan L Ameres[2,4,5], Gregor Obernosterer[3,5], Christoph A Gebeshuber[3], Renée Schroeder[2], Javier Martinez[3] & Ivo L Hofacker[1]

Small-interfering RNAs (siRNAs) assemble into RISC, the RNA-induced silencing complex, which cleaves complementary mRNAs. Despite their fluctuating efficacy, siRNAs are widely used to assess gene function. Although this limitation could be ascribed, in part, to variations in the assembly and activation of RISC, downstream events in the RNA interference (RNAi) pathway, such as target site accessibility, have so far not been investigated extensively. In this study we present a comprehensive analysis of target RNA structure effects on RNAi by computing the accessibility of the target site for interaction with the siRNA. Based on our observations, we developed a novel siRNA design tool, RNA*xs*, by combining known siRNA functionality criteria with target site accessibility. We calibrated our method on two data sets comprising 573 siRNAs for 38 genes, and tested it on an independent set of 360 siRNAs targeting four additional genes. Overall, RNA*xs* proves to be a robust siRNA selection tool that substantially improves the prediction of highly efficient siRNAs.

RISC executes the endonucleolytic cleavage of mRNAs in the final step of the RNA interference (RNAi) pathway[1,2], a process that is restricted by the accessibility of the target site both *in vitro* and under intracellular conditions[3]. To (i) comprehensively analyze target RNA effects on RISC function *in silico* and (ii) investigate the potential role of target site accessibility as an siRNA design criterion, we used the RNAplfold program[4]. RNAplfold allows fast computation of local base-pair probabilities and accessibilities within mRNA transcripts by sliding a short RNA folding window of a certain length over the longer mRNA sequence (**Fig. 1a**). More precisely, in our method, a window of size *W* is moved along the mRNA sequence, and the partition function for all local structures within the window is computed under the constraint that base pairing is allowed only between positions separated by, at the most, *L* nucleotides. RNAplfold assesses the accessibility of a region of length *u* by computing the probability that this stretch is unpaired in thermodynamic equilibrium and averages this accessibility over all windows containing the region. The sliding window approach offers several advantages. First, the procedure is very fast, as it can scan the whole 78 Mb of the human mRNAs in 12 h. Second, by averaging over many windows we avoid having to choose an arbitrary boundary surrounding the binding site. Moreover, we consider the complete Boltzmann ensemble of all possible structures rather than relying on a single predicted minimum free energy structure. Finally, we compute the exact accessibility rather than estimate it from a Boltzmann weighted sample of structures as is done in *S*fold[5], which may become inaccurate for longer regions with very low accessibility or longer length *u*.

To assess whether target site accessibility, as computed by RNAplfold, can be used to discriminate between functional and nonfunctional siRNAs and to determine the optimal parameters for *W*, *L* and *u*, we used two independent siRNA data sets of measured siRNA efficiencies. Data set 1 was composed of 2,433 siRNAs targeting the 3′ untranslated regions (UTRs) of 34 different genes[6], whereas data set 2 contained 294 siRNAs that were targeted against arbitrary regions of the coding sequences of the human genes *MAP2K1*, *GAPDH*, *PPIB* and *LMNA* and whose knockdown efficiencies were verified by analyzing mRNA as well as protein levels (**Supplementary Table 1** online). Data set 1 uses a so-called 'normalized inhibitory activity' to quantify target repression[6], whereas data set 2 uses the experimentally measured mRNA repression efficiency. Note that we refer to these data sets as the 'complete' data sets.

Many of the siRNAs in these data sets showed some, but relatively weak, repression efficiencies and were therefore not used in the training sets. From the initial data sets we generated reduced subsets of 474 and 99 siRNAs for data set 1 and 2, respectively, by removing siRNAs with intermediate silencing efficiencies, leaving only those that could be clearly assigned as functional or nonfunctional. The number of mRNAs targeted in the reduced data sets remained unchanged (see Methods for grouping). Target site accessibilities were then computed for different averaging window sizes *W*, maximum base pair spans *L* and lengths of the unpaired region *u*. To test for significant separation of functional and nonfunctional siRNAs in the two data sets, we performed a Wilcoxon rank sum test comparing the distributions of functional and nonfunctional siRNAs. We found that the silencing efficiency correlated significantly (minimum *P*-value of 2e-12 and 4e-4

[1]Institute of Theoretical Biochemistry (TBI), University of Vienna, Währingerstraße 17, 1090 Vienna, Austria. [2]Max F. Perutz Laboratories (MFPL), University of Vienna, Dr. Bohr-Gasse 9/5, 1030 Vienna, Austria. [3]Institute of Molecular Biotechnology (IMBA), Austrian Academy of Sciences, Dr. Bohr-Gasse 3, 1030 Vienna, Austria. [4]Present address: Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605-2324, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to I.L.H. (ivo@tbi.univie.ac.at) or J.M. (javier.martinez@imba.oeaw.ac.at).

Received 21 December 2007; accepted 7 April 2008; published online 27 April 2008; doi:10.1038/nbt1404

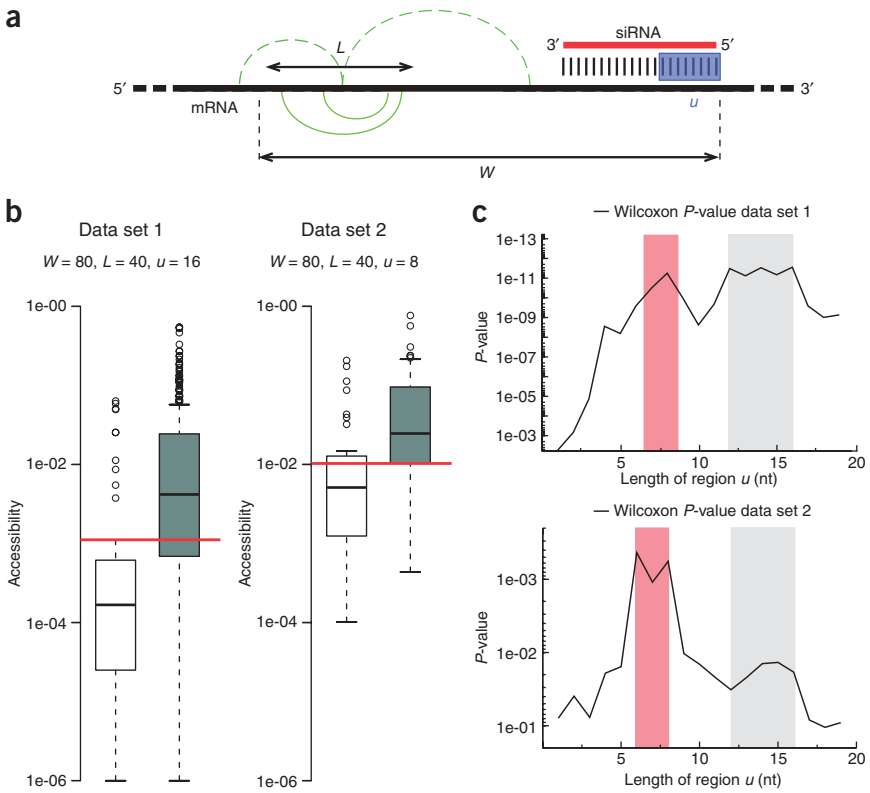

Figure 1 Application of RNAplfold to separate functional from nonfunctional siRNAs. (**a**) The RNA is folded locally in a sliding window approach (window size *W*). Within *W*, base pairing is restricted to a maximum distance *L*. *u* represents the stretch of consecutive nts within a siRNA target site starting at its 3′ end for which the accessibility is computed. Green lines represent possible base pairs. Interactions outside the span size of *L* or the flanking window *W* are not allowed (dotted green lines). (**b**) Box-plot diagram comparing the accessibility of functional and nonfunctional siRNAs. Two large data sets (data set 1 consisting of 474 siRNAs and data set 2 of 99 siRNAs) were used to determine optimal folding parameters for RNAplfold (for further detail see **Supplementary Fig. 1**). Data sets were divided into functional siRNAs (dark gray boxes, >0.900 repression score for data set 1 and repression >75% for data set 2) and nonfunctional siRNAs (white boxes, <0.354 repression score for data set 1 and repression <25% for data set 2). The quartiles are represented by the edges of the rectangles, which contain 50% of the data; black horizontal lines within the boxes depict medians. The circles represent outliers and dotted lines show the s.d. The Wilcoxon *P*-value is 2e-12 for data set 1 and 5e-4 for data set 2. Cutoffs for the accessibility to discriminate functional and nonfunctional siRNAs were set at 0.001002 for data set 1 and 0.01157 for data set 2 (red horizontal line). The parameters *W*, *L* and *u* are indicated. (**c**) Accessibility distributions of functional and nonfunctional siRNAs are best differentiated for a length of 8 and/or 16 nts (according to *P*-values). *P*-values were determined from a Wilcoxon test and are plotted against the length of the analyzed region starting at the 3′ end of the target site.

for data set 1 and data set 2, respectively) with target site accessibility over a wide range of parameters analyzed, with the most significant separation resulting from 80 nucleotides (nts) and 40 nts for *W* and *L*, respectively (**Fig. 1b** and **Supplementary Fig. 1** online). These values may seem small as they exclude any structure with base-pairs that span more than 40 nts. However, it is clear that actively translated coding regions are largely devoid of long range structures, because these structures are destroyed by the passing ribosome and are slow to re-form[7]. Moreover, it is well known that long range structures are much less accurately predicted[8]. Hence, a local structure approach may be more suitable than global mRNA structure prediction programs[5].

When varying the length *u* of the unpaired region, we observed two parameter ranges with especially good separation (**Fig. 1c**). The first region with a *P*-value peak measures the accessibility of the 6–8 nucleotides starting at the 3′ end of the target site and, therefore, corresponds to the so-called seed region. This is in agreement with previous observations that the 5′-seed region of both siRNAs and microRNAs is the major determinant for RISC-mediated target recognition[3,9–12]. Furthermore, a second region with a *P*-value peak was observed for *u* values of 12–16, reminiscent of biochemical data showing that accessibility of the first 16 nts within the target site is required for highly efficient RISC-mediated cleavage[3]. We also tested the stability of our parameter estimation by determining optimal parameters on subsets of the training data sets. Resampling always resulted in a *u* value of 8 and 16. The s.d. for *W* and *L* were 10.22 and 7.4 for data set 1 and 10.3 and 27.3 for data set 2, respectively. Although we observed both peaks in each of the data sets, the global optimum was $u = 16$ for data set 1 and $u = 8$ for data set 2. It is not clear whether this difference might be related to the fact that siRNAs target 3′ UTRs for data set 1 and coding regions for data set 2. In general, varying the length of the unpaired region *u* resulted in stronger effects on the separation of functional and nonfunctional siRNAs than varying *W* and *L* (data not shown). We were not able to detect any additional improvements in the separation of functional and nonfunctional siRNAs by also analyzing the energy of siRNA–target RNA duplexes. Presumably, there were no further improvements because perfect complementarity between siRNAs and their target sites generally implies high duplex energies. However, one cannot exclude such a correlation for siRNA off-target effects or microRNA-mediated gene repression, both of which rely on imperfect base pairing to their target sites[10,12,13].

Because it was previously claimed that regions of low G/C content coincide with efficient siRNA silencing[14] and because accessibility correlates with G/C content, we separated the data sets into five G/C content classes—35–45%, 40–50%, 45–55%, 50–60% and 55–65% G/C content—and analyzed the impact of accessibility for each class. We noticed that the distinction between functional and nonfunctional siRNAs remained strong over the whole range of G/C window sizes (**Supplementary Fig. 2** online). Furthermore, we found, that G/C content is a poorer predictor of siRNA efficiency than accessibility is. For data set 1, we noted that highly efficient siRNAs targeted regions of higher G/C content (on average 58%) whereas nonfunctional siRNAs had an average G/C content of 42%.

We also used the complete data set 1 (consisting of 2,433 siRNAs) to gain insight into the correlation between target site accessibility and siRNA repression efficiency. To reduce noise caused by the 30% error

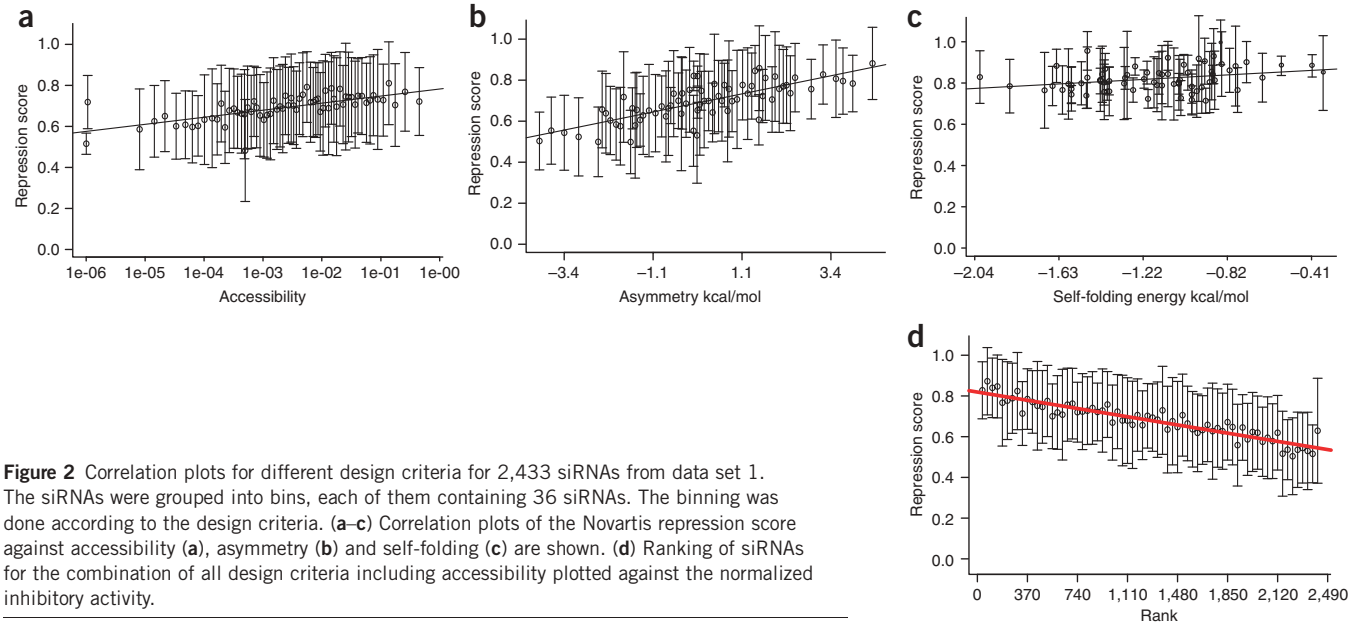

Figure 2 Correlation plots for different design criteria for 2,433 siRNAs from data set 1. The siRNAs were grouped into bins, each of them containing 36 siRNAs. The binning was done according to the design criteria. (**a**–**c**) Correlation plots of the Novartis repression score against accessibility (**a**), asymmetry (**b**) and self-folding (**c**) are shown. (**d**) Ranking of siRNAs for the combination of all design criteria including accessibility plotted against the normalized inhibitory activity.

in measured mRNA levels[6], we binned the data in groups of 36 siRNAs according to their accessibility and plotted the mean repression score against the mean accessibility for each bin. Despite the large variance, target accessibility clearly correlates with the repression score over a wide range of accessibilities (from 10^{-5} to 10^{-1}) (see **Fig. 2a**).

To assess the relevance of accessibility for the design of efficient siRNAs, we compared six commonly used criteria. Two criteria are purely sequence based ("U at position 10" and "a base other than G at position 13")[14]; two criteria describe the asymmetry of the siRNA duplex responsible for strand selection by looking at either the type of base pairs or the interaction energy of the last four base pairs[14–16]; and the final two look at the tendency for self-folding of the siRNA (which refers to the level of self-complementarity), using either its total folding energy (self-folding) or the number of unpaired nucleotides at the ends of the siRNA (free-end)[17]. All parameters were optimized on both reduced training data sets in a way similar to the accessibility criterion (see Methods). We noticed that the asymmetry resulted in a better correlation with the measured repression than the accessibility criterion (Pearson correlation coefficient of 0.48 and 0.23 for asymmetry and accessibility, respectively) (**Fig. 2b**). The other design parameters, free-end or self-folding, performed worse (**Fig. 2c**). The stronger effect of asymmetry could reflect the fact that selection of the siRNA strand occurs during RISC assembly, an event that takes place upstream of target site accessibility. We then designed simple filters for all the nonsequence–based criteria by defining a threshold. These thresholds were chosen in a conservative manner so that ≥75% of the functional siRNAs were retained from the complete data sets 1 and 2 (see Methods and **Supplementary Fig. 3** online). The performance of each filter was assessed on the complete data set 1 by applying a Wilcoxon test on the distribution of normalized inhibitory activities. We found that the two best single design criteria were accessibility and asymmetry with $P = 8.5\text{e-}8$ and $P < 1\text{e-}16$, respectively (see **Supplementary Table 2** online for a detailed overview). In addition, all siRNAs were binned in five functionality classes depending on their inhibitory efficiencies (inhibition smaller than 0.5, <F0.5; inhibition of at least 0.5, ≥F0.5; at least 0.8, ≥F0.8; at least 0.9, ≥F0.9; and at least 1.1, ≥F1.1). Even without any rational design many of the random siRNAs were functional with 82.3% inducing more than 0.5 inhibition (≥F0.5), 31.4% more than 0.8 (≥F0.8), 15.1% more than 0.9 (≥F0.9) and 2% more than 1.1 (≥F1.1). 'Random siRNAs' refers to siRNAs that were randomly chosen without any rational design. Free-end and self-folding performed slightly better than random, whereas the two sequence-based criteria did not result in any significant improvement, with "U at position 10" performing even worse than random (**Supplementary Table 2**). Therefore, the sequence rules were not considered further. From these analyses it is clear that accessibility alone, just like any other descriptor assessed above, is not sufficient to reliably predict siRNA efficacy. Because most current siRNA design methods neglect the effects of target site accessibility, we investigated whether the addition of accessibility to the three most effective conventional design criteria—asymmetry, free-end and self-folding—leads to a superior design for siRNAs. The combinations of asymmetry, self-folding and free-end led to an increase over random of 16.2%, 9.9% and 5.5% in ≥F0.8, ≥F0.9 and ≥F1.1. The addition of accessibility led to a further improvement in all functionality classes; specifically the fraction of siRNAs in the ≥F0.5, ≥F0.8, ≥F0.9 and ≥F1.1 classes increased by 3.4%, 3.9%, 4.2% and 2.1%, respectively (**Supplementary Table 2**). Especially the fraction of siRNAs in the ≥F0.9 (14.2%) and ≥F1.1 (7.6%) classes was doubled compared to random, demonstrating that the accessibility criterion boosts the fraction of very potent siRNAs. For a typical mRNA, ~25% of the sequence positions will pass the combination of all four filter criteria. Thus, the resulting list is usually long enough to choose specific siRNAs from the pool, an important feature, for example, for silencing specific gene splice variants or targeting short exons.

In addition to filtering, we introduced a ranking of the remaining siRNA candidates according to their overall performance in all four criteria. Because different selection criteria recapitulate distinct stages in the RNAi pathway, poor performance in one descriptor can most likely not be compensated for by good values in another. Rather than constructing a combined score, we therefore used a hierarchical sorting that emphasizes the least favorable criterion for each siRNA. The distribution of repression versus overall rank for data set 1 can be seen in **Figure 2d** and shows that, even among the siRNAs passing the

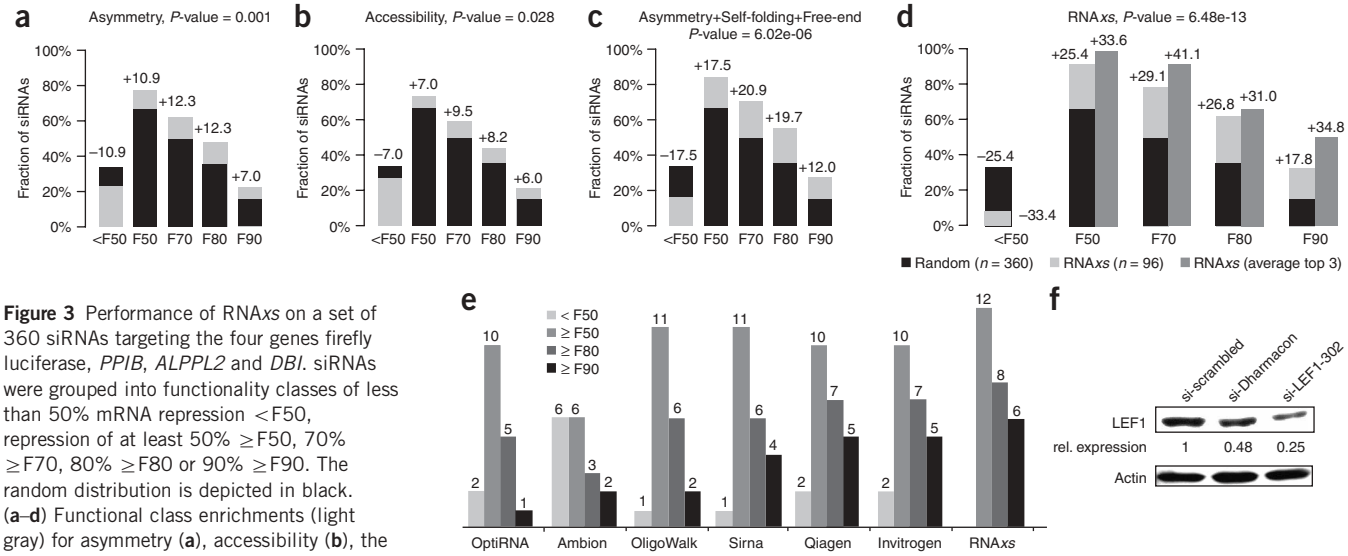

Figure 3 Performance of RNA*xs* on a set of 360 siRNAs targeting the four genes firefly luciferase, *PPIB*, *ALPPL2* and *DBI*. siRNAs were grouped into functionality classes of less than 50% mRNA repression <F50, repression of at least 50% ≥F50, 70% ≥F70, 80% ≥F80 or 90% ≥F90. The random distribution is depicted in black. (**a**–**d**) Functional class enrichments (light gray) for asymmetry (**a**), accessibility (**b**), the combination of asymmetry with self-folding plus free-end (**c**) and all parameters including accessibility (RNA*xs*) (**d**). The three top-ranked siRNAs are all contained in ≥F50 (dark gray). (**e**) Comparison of RNA*xs* to other design tools. OptiRNA[20], Ambion (siRNA Target Finder), Qiagen (siRNA Design Tool), Invitrogen (Block-iT RNAi Designer), oligowalk[21] and Sirna (using total score threshold; score > 12) were compared to RNA*xs* for the four functional classes (<F50, ≥F50, ≥F80, ≥F90). All tools were used with default parameters using the available web servers. For each tool, the repression efficiency of the three best-ranked siRNAs was assessed. RNA*xs* performed better than the other design tools for all functional classes. (**f**) Western blot analysis of extracts prepared from EpH4 cells, transiently transfected with scrambled siRNA, Dharmacon mmLEF1 SMARTpool (a combination of four siRNAs) or the single top-ranked siRNA designed with RNA*xs*. Relative LEF1 expression levels are indicated. Actin protein levels show equal loading. Full-length blots/gels are presented in **Supplementary Figure 7** online.

filters, the top-ranked candidates performed particularly well. The filtering and ranking described above were combined in a user-friendly siRNA design tool, called RNA*xs*, available as a web service at http://rna.tbi.univie.ac.at/cgi-bin/RNAxs. RNA*xs* returns a ranked list of all siRNA candidates that pass the thresholds for each design criteria, including their performance on each of the design criteria, as well as graphical accessibility plots for a user defined number of candidates. The user can change all parameters and thresholds. In addition, RNA*xs* enables entry of user-specific sequence constraints. Subsequently, individual sequences can be submitted to a BLAST search to detect possible off-target effects.

To test the performance of RNA*xs* in comparison to other methods, we used a third data set. This data set consisted of 360 siRNAs and was independent from the two data sets used to derive optimal design parameters and thresholds. Every second position of an arbitrarily chosen 198-nt stretch within firefly luciferase, human *PPIB*, *ALPPL2* and every position of a 108-nt stretch within *DBI* were targeted by an individual siRNA (**Supplementary Table 3** online). **Supplementary Figure 4** online shows the measured repression efficiency and target site accessibility plotted against the positions within the mRNA sequences for all four genes. The performance of this validation set confirmed our previous observations, namely that accessibility and asymmetry are the best single design criteria (**Fig. 3a**,**b**), whereas free-end and self-folding resulted only in a marginal improvement (**Supplementary Table 4** online). The combination of the three traditionally used criteria resulted in a significant ($P = 6e\text{-}6$) enrichment of effective siRNAs (**Fig. 3c**). The addition of accessibility to the three traditional criteria resulted in the best performance among all combinations (**Fig. 3d**). The enrichment for functional siRNAs was even better than in data set 1 (**Supplementary Table 4**). In contrast, using G/C content in addition to the three traditional criteria resulted in only a slight improvement (**Supplementary Table 4**). Note, that the *P*-values shown in **Figure 3** are not as high as for data set 1 simply because of the smaller number of siRNAs. On average, >90% of the rationally selected siRNAs were functional and almost every third siRNA reduced gene expression by >90%. Moreover, we looked specifically at the three top-ranked siRNAs for firefly luciferase, *PPIB*, *ALPPL2* and *DBI* (less amenable to silencing compared to the other three genes) and found all RNA*xs*-predicted siRNAs to be functional, half of them reducing gene silencing by >90% (see **Fig. 3d** and **Supplementary Fig. 5** online for a more detailed representation).

We compared RNA*xs* to six existing methods: three commercial siRNA selection tools and a machine-learning method, none of which considers target accessibility; oligowalk, a machine-learning method that considers target accessibility; and Sirna[18], which assesses target site accessibility as computed by *S*fold in combination with duplex stability. Because some of these methods return only a few siRNA candidates, we limited the comparison to the top three siRNAs predicted by each tool for each of the four genes. We compared the predicted siRNAs with the measured silencing efficiencies by sorting them into different functionality classes of less than 50% (<F50), more than 50% (≥F50), more than 80% (≥F80), and more than 90% (≥F90) mRNA repression. RNA*xs* was the only tool where all predicted siRNAs had a measured repression efficiency of ≥F50; in fact, it outperformed all other programs in each of the functionality classes (**Fig. 3e**). We furthermore performed a gene knockdown experiment of the murine lymphoid enhancer-binding factor 1 (LEF1) protein by using the single top-ranked siRNA from RNA*xs* as well as a commercial siRNA pool and measured the resulting protein levels. The pool, which consisted of a combination of four rationally chosen siRNAs, resulted in ~50% knockdown of LEF1, whereas the single siRNA designed with RNA*xs* resulted in 75% protein knockdown efficiency (**Fig. 3f**). When the respective target sites were analyzed in more detail, we found that only one of the four

siRNAs of the pool would pass the RNA*xs* filters, whereas the other three would be rejected due to accessibility (two out of three) or asymmetry (one out of three) (**Supplementary Fig. 6** online).

In summary, we have shown that the accessibility criterion as computed by the RNAplfold program allows incorporation of the influence of target mRNA structure in the siRNA design process. When combined with previously reported criteria for the design of potent siRNAs, the computation of target site accessibility significantly improves the prediction of effective siRNAs. Furthermore, we present an siRNA design tool called RNA*xs* that implements these criteria in the form of filters and a ranking procedure. Our approach is based exclusively on biologically comprehensible design criteria describing distinct stages in the RNAi pathway, provides adjustable parameters and supplies the user with a ranked list that contains all siRNAs that fulfill our criteria. This allows the selection of siRNA with special properties, such as those targeting specific splice variants or those targeting homologous genes. In summary, RNA*xs* represents an siRNA design tool that results in a substantial increase of highly effective siRNAs.

METHODS

SiRNA data sets, grouping and synthesis. Data set 1 was composed of 2,433 siRNAs targeting the 3′ UTR sequences of 34 different genes. This data set was taken from ref. 6. Data set 2 contained 294 newly synthesized siRNAs that were targeted against arbitrary regions of the coding sequences of the human genes *MAP2K1*, *GAPDH*, *PPIB* and *LMNA* and whose knockdown efficiencies were verified by analyzing mRNA as well as protein levels. The full list of siRNA sequences, position of targets and repression efficiencies are reported in **Supplementary Tables 1** and **3**. Note that the two data sets use different scores to quantify for the siRNA repression efficiency. For data set 1 (ref. 6) the "normalized inhibitory activity" is used, which is obtained by scaling linearly the averaged raw inhibitory activities such that their positive control becomes 0.900 "normalized inhibitory activity" and their negative control becomes 0.354 "normalized inhibitory activity," the latter to ensure that no negative activity exists. In contrast to data set 1, data set 2 (siRNAs were synthesized by Dharmacon Inc.) uses the averaged measured mRNA repression levels.

From these complete data sets we generated subsets of 474 and 99 siRNAs for data sets 1 and 2, respectively, by removing siRNAs with intermediate silencing efficiencies. The functional group for data set 1 contained siRNA with a normalized inhibitory score >0.900, whereas the nonfunctional group contained only siRNAs with a score <0.354. Those thresholds correspond to the inhibitory score of the positive and negative control, respectively. For data set 2, the repression efficacies of the functional group were $>75\%$, whereas the nonfunctional group contained only repression $<25\%$.

The siRNAs used to evaluate the performance of RNA*xs* were designed to target every other position of the following regions on the following genes (198 nts for each gene): human *PPIB*, 374–570, RefSeq accession no. NM_000942; firefly luciferase, 1713–1909, RefSeq accession no. U47298 (pGL3, Promega); human *ALPPL2*, 261–457, RefSeq accession no. NM_031313.2 and as well as to target every position of the following region of human *DBI* (108 nts), 413–520, GenBank accession no. NM_020548.3.

All siRNAs for the knockdown of *MAP2K1*, *GAPDH*, *PPIB*, *LMNA*, firefly luciferase, human *ALPPL2* and human *DBI* are based on duplexes synthesized by Dharmacon Inc. as 21-mers with 3′-dTdT overhangs, as described previously[14]. See **Supplementary Table 3** for the full list of siRNAs, their target site positions as well as their respective averaged repression efficiencies.

The guide strand sequences of the four siRNAs of the Dharmacon SMARTpool (sequence information provided by Dharmacon Inc.) against murine lymphoid enhancer-binding factor 1 (LEF1) were as follows: 5′-CGAAGAG GAGGGCGACUUAdTdT-3′, 5′-CGUCAGAUGUCAACUCCAAdTdT-3′, 5′-GGAGUCGACUUCAGGUACAdTdT-3′ and 5′-AAUGAGAGCGAAUGUC GUAdTdT-3′ (see **Supplementary Fig. 6**). The RNA*xs* designed siRNA guide sequence was 5′-CUCUUUUUCUCCCCUCCUCCU-3′ and the passenger sequence was 5′-GAGGAGGGGAGAAAAAGAGAA-3′ (synthesized and de-protected by Sigma-Proligo). All data sets can also be accessed at http://www.tbi.univie.ac.at/~htafer/RNAxs.

siRNA nomenclature. All siRNA duplexes are referred to by sense strand. The siRNA functionality nomenclature was as described previously[14]. Briefly, the first nt of the 5′end of the sense strand is position 1, which corresponds to position 19 of the anti-sense strand. The following system of nomenclature was used to compare and report siRNA functionality: 'F' followed by the degree of minimal knockdown in percent. For example, F50 signifies at least 50% knockdown, F80 means at least 80% knockdown, and so forth. All siRNAs below F50 were considered to be nonfunctional.

For data set 1, the nomenclature for the functionality groups was based on the normalized inhibitory activity, which was previously described[6].

Cell culture, transfection and quantification of gene knockdown. Cell culture, transfections and quantifications were performed as described previously[14]. For the RNAi knockdown experiment of murine LEF1, EpH4 mouse mammary epithelial cells were transiently transfected with 25 nM siRNA (either Dharmacon SMARTpool or RNA*xs*-designed siRNA) using DharmaFECT 1 transfection reagent (Dharmacon Inc.) according to the instructions of the manufacturer. Cells were washed with PBS 48 h after transfection and lysed on ice with Nonidet P-40 lysis buffer centrifuged at 12,000*g* for 10 min at 4 °C. To quantify LEF1 gene knockdown efficacy, freshly prepared lysate supernatants were subjected to 10% SDS-PAGE and immunoblotted as described previously[19], using mouse-anti LEF-1 (Upstate, clone 4H2) and rabbit-anti beta-actin (Sigma-Aldrich, clone AC-74) antibodies.

Local folding using RNAplfold and parameter selection. Accessibility parameters for mRNAs were determined using the RNAplfold program, which implements a local folding algorithm to compute base-pairing probabilities and accessibility in a scanning window approach (window size W). The locality of the structure is determined by L, which specifies the maximal distance between two pairing positions. Accessibility is measured as the probability that a region of predefined length u is free of base pairing in thermodynamic equilibrium. Base-pairing probabilities (restricted by L) and accessibility (of the region u) are averaged over all windows W ($W > L$) that contain the stretch u. To assess optimal folding parameters for the partition of functional and nonfunctional siRNAs, we performed a grid search by varying the window size W from 20 to 200 nts in a step size of 20 nts, the maximal binding lengths to $L = 1/4W$, $L = 1/2W$, $L = 3/4W$ and W-5 and the size of the region for which the accessibility was computed from 1 to 19 nts in steps of 1 nt. A Wilcoxon test (a nonparametric test for assessing whether two samples of observations come from the same distribution) was applied on the two sets of functional and nonfunctional siRNAs for all W, L, u, triplets (**Supplementary Fig. 1**). The parameter set ($W = 80$ nts, $L = 40$ nts, $u = 16$ nts for data set 1 or 8 nts for data set 2) gave the overall best P-values as determined by a Wilcoxon test and were therefore considered as the optimal parameters for siRNA prediction and subsequently used in *RNAxs*. Note, that for an siRNA to be selected by RNA*xs* it is required that both accessibility scores ($u = 8$ and $u = 16$) lie above thresholds.

Accessibility thresholds. Based on the parameters W, L, u that gave the best separation (and the least number of false positives) on functional and nonfunctional siRNAs cutoffs for choosing potent siRNAs were selected (see also red lines, **Fig. 1b**). The optimal size for W was 80 nts, the maximal allowed base-pairing distance L determined as 40 nts and the accessibility thresholds for $u = 8$ (data set 2) and $u = 16$ (data set 1) set at 0.01157 and 0.001002, respectively. Overall, for data set 2, where the optimal u was 8 nts, the threshold gives a true-positive rate of 75% and a false-positive rate of 27%, whereas with data set 1, where the best u was 16 nts, we obtained a true-positive rate of 70% and false-positive rate of 20%.

siRNA design criteria. We investigated the significance of several biologically comprehensible and previously published siRNA design criteria on data set 1 and 2. Namely, siRNA strand selection, referred to as asymmetry, self-folding, free-end and accessibility as computed by RNAplfold were analyzed. Asymmetry was assessed through a sequence and an energy-based rule. For the

sequence rule, the number of G/Cs of the first two nts at the 5′ end of the sense strand was subtracted from the number of G/Cs of the first two positions at the 5′ end of the anti-sense strand. The threshold on this parameter was set to 0 (same number of G/Cs on both ends). For the energy-based rule, the duplex energy for the first three nucleotides on both 5′ ends was computed and the difference calculated. A lower thermodynamic duplex stability at the 5′ end of the anti-sense strand favors its selection when compared to the sense strand. The threshold was set to –0.39 kcal/mol. Asymmetry specifies strand selection and, therefore, acts upstream of any target interaction. The design criteria self-folding and free-end describe the influence of the guide strand structure on gene silencing. Self-folding is measured as the minimal free folding-energy of the guide strand, as computed using RNAfold, and the respective threshold was set to –2 kcal/mol. The free-end parameter was assessed as the number of paired nucleotides among the first four nts at the 5′ end and 3′ end of the guide strand. Those thresholds were conservatively set so that at least 75% of the functional siRNAs of the functional siRNAs in both data sets were kept (**Supplementary Fig. 3**). It should be noted that the design criteria other than accessibility were optimized in a similar manner; for example, to determine the optimal number of terminal nucleotides for asymmetry and free-end, the number of terminal nucleotides was varied from 1 to 5 in a step-size of 1 nt.

Statistical analysis. *P*-values were calculated using the nonparametric Wilcoxon rank sum test using the freely available statistical computing software R.

Accessibility and repression plots. All plots were generated using the xmgrace and R 2D plotting tools.

RNA*xs*. RNA*xs* combines RNAplfold and previously reported siRNA design criteria[14–17]. RNA*xs* reads RNA sequences in FASTA format and offers the user the possibility to modify predefined thresholds for asymmetry, self-folding, free-end and target site accessibility. The user can choose from a number of predicted siRNAs, which are ranked according to their overall performance over all design criteria. RNA*xs* returns accessibility plots for the sequence(s) analyzed, which can be inspected visually. A zoomed plot for each predicted siRNAs of 100 nts surrounding the target site is provided. A table, containing the scores for the different siRNA design parameters for each siRNA is returned. SiRNAs can be ranked by either choosing one of the design parameters or by applying the 'multiple criteria ranking' option. In the latter case we first rank all siRNAs by each of the six design parameters (two for accessibility, two for asymmetry, self-folding and free-end) separately, thus assigning six rank numbers to each siRNA. The overall sorting is then a hierarchical sort using the worst rank (WR) as the primary sorting key. Therefore, the best predicted siRNAs are the ones having the smallest WR. RNA*xs* is available as a web service under the address http://rna.tbi.univie.ac.at/cgi-bin/RNAxs. The source code can be accessed at: http://www.tbi.univie.ac.at/~htafer/RNAxs.

G/C-content analysis. The training data sets were divided into groups based on their G/C-content (35–45, 40–50, 45–55, 50–60, 55–65). Note, that there was not a sufficient number of siRNAs with a G/C content of $<35\%$ and $>65\%$ to reliably compute *P*-values. RNAplfold was used to calculate the accessibility for the two functionality groups $<$F30 and $\geq$F80 in the five G/C-content subgroups (**Supplementary Fig. 2**). RNAplfold was used with the previously defined optimal folding parameters ($W = 80$, $L = 40$ and $u = 16$).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We would like to thank A. Khvorova for critical discussions during the project as well as Dharmacon Inc. for providing siRNA knockdown data. H.T. is supported by Siemens and the Wiener Wissenschafts-, Technologie-, und Forschungsfonds. S.L.A. is funded by the Austrian Science Fund FWF through WK001. J.M. is a Junior Group Leader at IMBA, the Institute of Molecular Biotechnology supported by the Austrian Academy of Sciences. Funding by the Austrian Government's GEN-AU is acknowledged by R.S., J.M. and I.L.H. We thank the members of the Hofacker, Schroeder and Martinez labs for encouragement, helpful discussions and comments on the manuscript.

AUTHOR CONTRIBUTIONS

H.T., S.L.A. and G.O. initiated research, designed and performed the experiments and analyzed data. I.L.H. designed algorithms, supervised implementation and analysis. C.A.G. performed western blot analysis. R.S. supervised experimental work and analysis. J.M. supervised experimental work and analysis and procured access to the Dharmacon data. All authors contributed to the writing of the manuscript.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Tomari, Y. & Zamore, P.D. Perspective: machines for RNAi. *Genes Dev.* **19**, 517–529 (2005).
2. Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**, 343–349 (2004).
3. Ameres, S.L., Martinez, J. & Schroeder, R. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell* **130**, 101–112 (2007).
4. Bernhart, S.H., Hofacker, I.L. & Stadler, P.F. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22**, 614–615 (2006).
5. Ding, Y. & Lawrence, C.E. Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic Acids Res.* **29**, 1034–1046 (2001).
6. Huesken, D. *et al.* Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.* **23**, 995–1001 (2005).
7. Takyar, S., Hickerson, R.P. & Noller, H.F. mRNA helicase activity of the ribosome. *Cell* **120**, 49–58 (2005).
8. Doshi, K.J., Cannone, J.J., Cobaugh, C.W. & Gutell, R.R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 105 (2004).
9. Haley, B. & Zamore, P.D. Kinetic analysis of the RNAi enzyme complex. *Nat. Struct. Mol. Biol.* **11**, 599–606 (2004).
10. Brennecke, J., Stark, A., Russell, R.B. & Cohen, S.M. Principles of microRNA-target recognition. *PLoS Biol.* **3**, e85 (2005).
11. Doench, J.G. & Sharp, P.A. Specificity of microRNA target selection in translational repression. *Genes Dev.* **18**, 504–511 (2004).
12. Jackson, A.L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.* **21**, 635–637 (2003).
13. Lai, E.C. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30**, 363–364 (2002).
14. Reynolds, A. *et al.* Rational siRNA design for RNA interference. *Nat. Biotechnol.* **22**, 326–330 (2004).
15. Schwarz, D.S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208 (2003).
16. Khvorova, A., Reynolds, A. & Jayasena, S.D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216 (2003).
17. Patzel, V. *et al.* Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency. *Nat. Biotechnol.* **23**, 1440–1444 (2005).
18. Ding, Y. & Lawrence, C.E. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* **31**, 7280–7301 (2003).
19. Yu, Y. & Sato, J.D. MAP kinases, phosphatidylinositol 3-kinase, and p70 S6 kinase mediate the mitogenic response of human endothelial cells to vascular endothelial growth factor. *J. Cell. Physiol.* **178**, 235–246 (1999).
20. Ladunga, I. More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. *Nucleic Acids Res.* **35**, 433–440 (2007).
21. Lu, Z. & Mathews, D.H. Efficient siRNA selection using hybridization dynamic. *Nucleic Acids Res.* **36**, 640–647 (2008).