nature biotechnology

Design of a genome-wide siRNA library using an artificial neural network

Dieter Huesken[1,6], Joerg Lange[1,6], Craig Mickanin[2], Jan Weiler[1], Fred Asselbergs[1], Justin Warner[2], Brian Meloon[3,5], Sharon Engel[4], Avi Rosenberg[4], Dalia Cohen[2], Mark Labow[2], Mischa Reinhardt[1], François Natt[1] & Jonathan Hall[1]

The largest gene knock-down experiments performed to date have used multiple short interfering/short hairpin (si/sh)RNAs per gene[1–3]. To overcome this burden for design of a genome-wide siRNA library, we used the Stuttgart Neural Net Simulator to train algorithms on a data set of 2,182 randomly selected siRNAs targeted to 34 mRNA species, assayed through a high-throughput fluorescent reporter gene system. The algorithm, (BIOPREDsi), reliably predicted activity of 249 siRNAs of an independent test set (Pearson coefficient $r = 0.66$) and siRNAs targeting endogenous genes at mRNA and protein levels. Neural networks trained on a complementary 21-nucleotide (nt) guide sequence were superior to those trained on a 19-nt sequence. BIOPREDsi was used in the design of a genome-wide siRNA collection with two potent siRNAs per gene. When this collection of 50,000 siRNAs was used to identify genes involved in the cellular response to hypoxia, two of the most potent hits were the key hypoxia transcription factors HIF1A and ARNT.

Evidence suggests that motifs in the siRNA and/or the target mRNA largely determines inhibitory activity of siRNAs[4]. Current tools created to recognize such motifs have been developed from studies on localized hybridization energy in parts of the siRNA duplex, on the occurrence of preferred nucleotides at specific positions and on H-bonding patterns. Analysis of a moderately large data set of functional micro (mi)RNA and siRNA guides has revealed that the 5′-ends are consistently rich in A/U[5,6]. Furthermore, uridine is often present at positions 1 and 17 of the guide strand, and an adenosine moiety at position 10 (ref. 7). Finally, strong internal structures, an absence of G/C stretches and a prevalence of certain nucleotides at positions 4, 7, 9, 14 and at the 3′-end are also characteristic of active siRNAs[8–10]. Although there is consensus on the importance of features at the 5′-terminus of the guide strand, there is little agreement on the other positions possibly because the data sets used are often too small to ensure statistical significance[11]. It is also likely that the complexity of motifs and other strand characteristics that promote the RNAi mechanism depend on more than just the single nucleotide composition. Spontaneous hydrolysis and ribonuclease A–induced hydrolysis is accelerated at pyrimidine-A dinucleotides[12], and influenced by neighboring nucleotides and hydrogen bonds[13]. Such motifs are difficult to detect by hypothesis-driven inspection of sequences in small data sets. Furthermore, as more motifs are discovered, it is more difficult to identify target-specific siRNAs that carry all motifs. Consequently, they need to be weighted, a process that is usually based upon arbitrarily selected factors[7].

Artificial neural networks (ANNs), often referred to as 'black boxes' because the parameters that they derive to approximate patterns cannot be easily analyzed[14], provide a powerful method of identifying highly complex traits in data sets. ANNs discover and work with large numbers of interrelated motifs developed through automated unbiased learning. They then combine them for accurate prediction using their own weighting systems. ANNs have been broadly applied in the biological sciences, for example, for predicting drug mechanisms[15] and 3D-protein structure[16] and also, to identify active antisense oligonucleotides (ASOs)[17,18]. The prediction quality and generalization[19] capabilities of an ANN of fixed size depend on a sufficiently large training set of directly comparable data points. However, as oligonucleotide activity is highly sensitive to several biological and experimental parameters (e.g., transfection efficiency, target metabolism, biology), direct comparison of siRNAs between assays or even across targets in the same cell line is usually difficult. We concluded that there was no published data set suitable for ANN training. We reasoned that an algorithm that predicts relative potencies of siRNAs targeted to an ectopically expressed reporter gene should function equally well on endogenously expressed genes, and that a large homogeneous data set for training ANNs could be best generated in a high-throughput reporter screening assay.

We modified a previously described reporter assay[20] using a plasmid coding for both a reporter gene (enhanced yellow fluorescent protein, eYFP) bearing target cDNA inserts in its 3′-untranslated region (UTR; that is, a fusion mRNA) and a reference gene (enhanced cyan fluorescent protein, eCFP) (see **Supplementary Fig. 1a** online).

[1]Novartis Institutes for BioMedical Research, Genome and Proteome Sciences, CH-4002 Basel, Switzerland. [2]Novartis Institutes for BioMedical Research, Inc., Genome and Proteome Sciences, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [3]Compugen USA, Inc., 7 Centre Drive Suite 9, Jamesburg, New Jersey 08831, USA. [4]Computational Life Sciences R&D, Compugen Ltd., 72 Pinchas Rosen St., Tel Aviv 69512, Israel. [5]Present address: Campbell & Company, Inc., 210 West Pennsylvania Ave., Suite 770, Towson, Maryland 21204, USA. [6]These authors contributed equally. Correspondence should be addressed to J.H. (jonathan.hall@Novartis.com).

Published online 17 July 2005; doi:10.1038/nbt1118

Hybridization of siRNAs to sites in the inserts attenuates eYFP protein levels only. Although modification of the 3′-UTR of an mRNA might be expected to affect its regulation, we have generally observed that levels of expressed protein vary three- to fourfold at most. Cotransfection of a plasmid and NAS 12842, a potent eYFP siRNA targeted to a common site in its 5′-UTR, resulted in a specific dose-dependent decrease in eYFP expression (**Supplementary Fig. 1b** online). Furthermore, having six mismatched nucleotides to eCFP, NAS 12,842 showed no effect on eCFP expression (data not shown). Moreover, of seven additional eYFP siRNAs with ≥85% identity to eCFP mRNA, five showed a near perfect specificity for eYFP (data not shown). To demonstrate a correlation between the potency profile of a set of siRNAs targeted to the reporter gene and the profile obtained from targeting the endogenously expressed mRNA, we designed and tested 37 siRNAs (see **Supplementary Table 1** online) against the following three cDNA target inserts: a 3′-UTR sequence from the *TC10* ras-like gene and coding regions from ubiquitin conjugating enzymes (E2s) *UBE2I* and *CDC34*. Inhibition of endogenous *TC10* mRNA expression was measured using quantitative reverse transcriptase PCR (Q-PCR), whereas inhibition of the E2s was assayed by western blot analysis.

Activities of 12 *TC10* siRNAs in the reporter assay varied broadly and less than half of them inhibited eYFP expression by ≥50% (**Fig. 1a**). The range of activities against the endogenously expressed target was narrower (**Fig. 1b**), even at higher siRNA concentrations and, also, if tested with a second set of primer/probes (data not shown). A standard Pearson correlation coefficient was calculated for the data sets (**Fig. 1c**): $r = 0.89$ ($P = 1.2 \times 10^{-4}$). The high correlation on this relatively small sample size is consistent with similar experiments using ASOs[20]. Of fourteen siRNAs targeted to *UBE2I*, five inhibited eYFP by ≥50% (**Fig. 1d**) and were among the most inhibitory oligonucleotides as measured by quantification of residual endogenous *UBE2I* protein (**Fig. 1e**). A Pearson correlation coefficient of $r = 0.83$ ($P = 2.3 \times 10^{-4}$) was obtained (**Fig. 1f**). Similarly, of eleven siRNAs targeting *CDC34* (**Fig. 1g**), less than half inhibited the

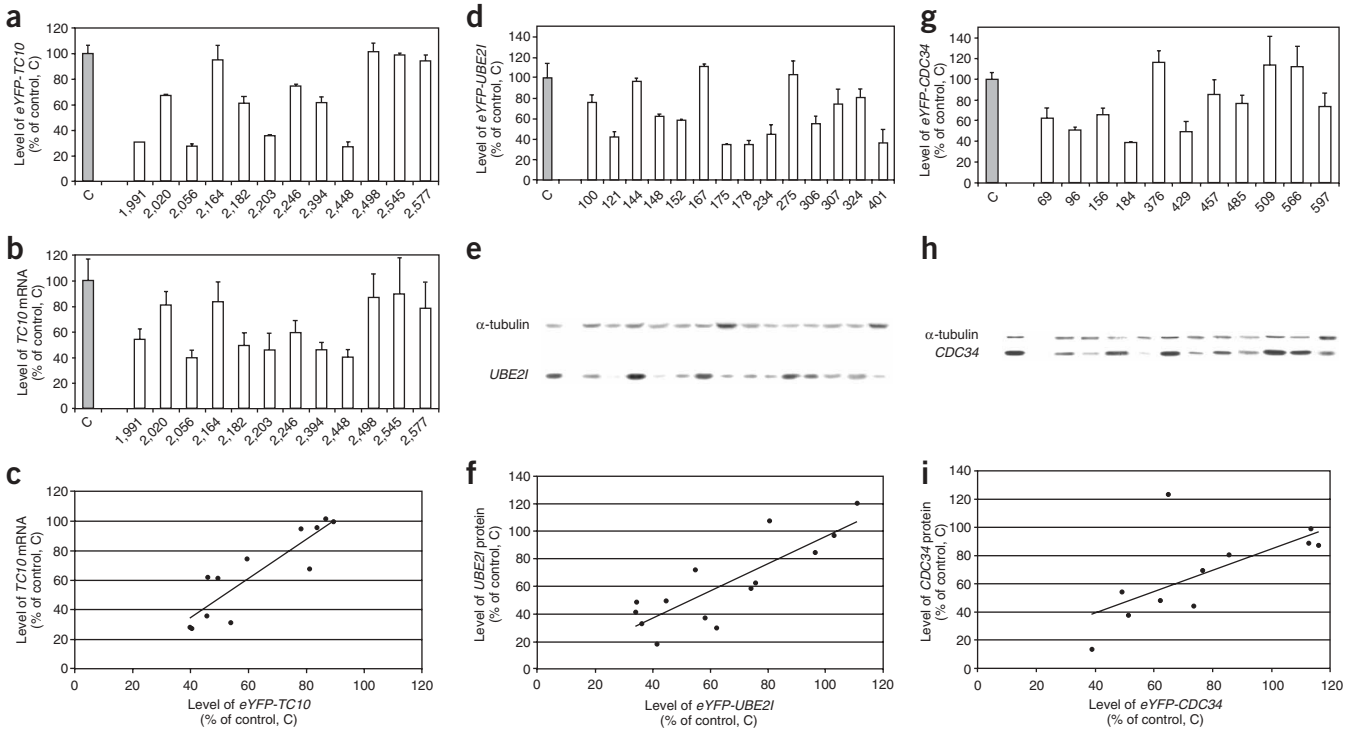

Figure 1 Thirty seven siRNAs were designed and tested against three human targets (for siRNA sequences, see **Supplementary Table 1**). Inhibitory activity was measured in the dual reporter assay and also against endogenously expressed genes at the mRNA or protein levels. siRNAs are identified by their position relative to the start codon. The reporter assay was expressly performed at suboptimal concentrations so as to obtain the broadest possible window of inhibitory activity. Protein levels were normalized to α-tubulin. (**a**) Downregulation of eYFP- *TC10* reporter gene fusion by twelve siRNAs targeted to various positions in the predicted 3′-UTR sequence of the *TC10* ras-like gene (NM_012249) (50 nM, 48 h) in H1299 compared to an unrelated negative control siRNA NAS 8549 (C), measured by eYFP fluorescence. (**b**) Downregulation of endogenous *TC10* mRNA in H1299 measured by real-time reverse-transcriptase PCR (Q-PCR: primer and Taqman probes positioned in the 3′-UTR) with siRNAs (10 nM, 48 h) compared to control (C). (**c**) Two-dimensional plot with linear regression line of siRNAs targeting endogenous *TC10* mRNA and a *TC10*-containing eYFP reporter construct: Pearson correlation coefficient: $r = 0.89$. (**d**) Inhibition of eYFP-*UBE2I* reporter gene fusion (H1299) with fourteen siRNAs (50 nM, 48h) targeted to various positions in the coding region of the human *UBE2I* gene (NM_003345), compared to control (C) measured by eYFP fluorescence. (**e**) Western blot analysis showing inhibition of endogenous *UBE2I* protein with the siRNAs (H1299, 25 nM, 48h) compared to control (C): α-tubulin was used to normalize *UBE2I* protein expression values for slight variations in protein gel loading. (**f**) Two-dimensional plot with linear regression line of siRNAs targeting endogenous *UBE2I* protein and eYFP-*UBE2I* reporter: Pearson correlation coefficient: $r = 0.83$. (**g**) Inhibition of eYFP- *CDC34* reporter gene fusion (H1299) with 11 siRNAs (50 nM, 48h) targeted to various positions of the *CDC34* coding region (NM_004359), compared to control (C), measured by eYFP fluorescence. (**h**) Western blot analysis showing inhibition of endogenous *CDC34* protein with the siRNAs (H1299, 25 nM, 48h) compared to control (C). (**i**) Two-dimensional plot with linear regression line of siRNAs targeting endogenous *CDC34* protein and eYFP-*CDC34* reporter: Pearson correlation coefficient: $r = 0.66$. The apparent outlier CDC34 siRNA of mediocre inhibitory activity in the reporter assay was stimulatory when assayed against endogenous CDC34 protein. In HeLa and KB-31 cells the same batch of siRNA behaved as expected (data not shown) and so we assume that this is a nonspecific effect in H1299.

reporter gene expression by ≥50%. With the exception of one siRNA, members of this sub-group were also among the most inhibitory at the protein level (**Fig. 1h**). A Pearson correlation coefficient of $r = 0.66$ ($P = 2.8 \times 10^{-2}$) was obtained for *CDC34* (**Fig. 1i**). The high correlation coefficients in these relatively small experiments (37 siRNAs) indicate that potency profiles of siRNAs against the reporter fusion mRNA and the corresponding endogenous gene are similar: a poor correlation here would have strongly implied that nucleotide sequence is of lesser importance to potency than, for example, mRNA abundance, half-life, processing or transport. It also implies that common sequence segments in endogenous and fusion RNAs have similar secondary/tertiary structures, or alternatively, that the RNAi mechanism can overcome RNA folding. We favor the former explanation, without discounting the latter, as analogous results were obtained for ASOs, which operate via a different mechanism. In summary, we concluded that the reporter assay was a suitable means to generate large homogeneous data sets for training ANNs.

ANNs were generated with data according to a flowchart (**Fig. 2a**). Thirty-four constructs with fully sequenced inserts derived from nineteen E2s, seven other human and eight rodent genes (see **Supplementary Table 2** online) yielded 27,000 nucleotides of total target sequence, which amounted to ~27,000 possible tiled siRNAs. Then, we screened 3,106 randomly selected siRNAs in which all possible di- and trinucleotides were frequently represented at each possible starting position, in H1299 cells in duplicate. Positive and negative controls for normalization enabled comparison of oligoribonucleotide activity across assays. eCFP served as a control for efficiency of plasmid transfection, eYFP-siRNA 12842 served to normalize for oligonucleotide transfection efficiency and an anti-pGL3 luciferase siRNA was used to control for nonspecific inhibition of eYFP. Data from 2,431 siRNA sequences (see **Supplementary Table 3** online) passed quality control filters and produced an approximate Gaussian distribution of potencies relative to positive and negative controls (**Fig. 2b**). As this data set is unique to our knowledge, we examined simple motifs in the 8% most active sequences, that is, the top 200 siRNAs, using stringent criteria for statistical significance. We studied each position of the guide sequence of potent siRNAs for unambiguous mononucleotide motifs present in relatively high excess over expected occurrence with significance $P \leq 0.05\%$. For each motif identified, we then examined the occurrence of nucleotides at the same position in the 200 least potent siRNAs. Although it is not given that a motif that contributes to the potency of an siRNA should automatically be absent in inactive siRNAs, consistency between active and inactive siRNAs for the occurrence of a motif at a given position should help to minimize false positives. All motifs with $P \leq 0.05\%$ are shown in **Supplementary Table 4** online. The most overrepresented ('dominant') and most highly statistically significant mono-nucleotide motifs in potent siRNAs were A and U at position 1, A at position 10 and U at position 2. The requirement for A/U at the 5′-guide terminus was discovered and rationalized previously[5,6], and A at position 10 corresponds to a previously characterized U-cleavage site[21]. The high occurrence of U at position 2 is new and does not reflect the need for weak affinity close to the 5′-end, as A was also not overrepresented at this position. It may induce a favored interaction with the RNA-induced silencing complex (RISC), or provide a reactive group important for the RISC mechanism. Other mono-nucleotides were overrepresented at positions 7 (U), 11 (U) and 19 (C) of active siRNAs, while being underrepresented in inactive siRNAs. It should also be noted that even higher levels of statistical significance and overrepresentation were observed for alternative mono-nucleotides

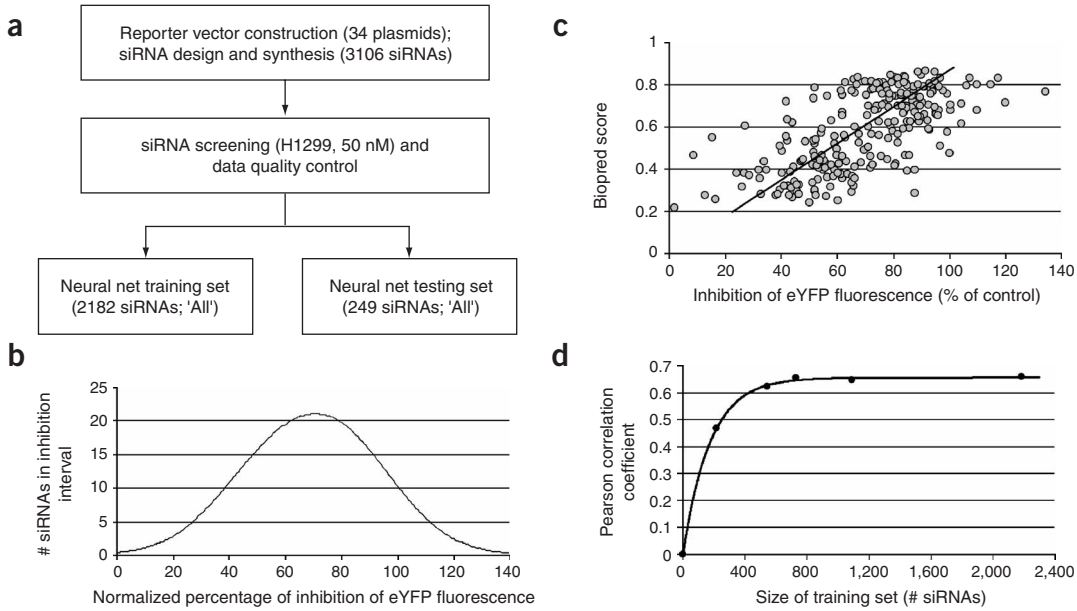

Figure 2 Management of data during the training and testing of artificial neural networks (ANN). (**a**) Experimental design for the generation and testing of ANNs. siRNAs were selected randomly across each insert, excluding the first and terminal 50 nucleotides, with an average overlap of 9 nucleotides, at an siRNA concentration of 50 nM; data were collected at two time points (**b**) Estimated density distribution of inhibitory activities from folding a Gaussian kernel of width 0.10 with the measures of 2,431 randomly selected siRNAs targeted to 34 reporter plasmids; sequences of plasmid inserts, and sequences of the siRNAs used in the investigation are given in **Supplementary Tables 2** and **3**, respectively. *Y*-axis normalized for correspondence to a 256-bin histogram. The *x*-axis is normalized with the positive control set to 90.0% inhibition and the negative control is set to 35.4% inhibition, such that the least active siRNA becomes 0%. (**c**) Plot of predicted inhibition of 249 siRNAs by the ANN against observed inhibitory activity obtained in the eYFP reporter assay (Pearson correlation coefficient $r = 0.66$). (**d**) Performance of ANNs on the test set 'All' as a function of size of the training set (218, 545, 727, 1,091 and 2,182 siRNAs).

Table 1 Pearson correlation coefficients of ANNs generated from data sets of various sizes and compositions

	Testing set (21-nt)			
Training set (21-nt)	All (249)	All human (198)	hE2 (139)	Rodent (51)
All (2,182)	0.66	0.63	0.63	0.77
All human (1,744)	0.65	0.61	0.62	0.76
Human E2s (1,229)	0.65	0.62	0.62	0.76
Rodent (438)	0.55	0.54	0.53	0.57
Random all (1,091)	0.65	0.62	0.61	0.75
Random all (727)	0.65	0.63	0.63	0.76
Random all (545)	0.62	0.60	0.60	0.70
Random all (218)	0.47	0.47	0.46	0.46
All-19	0.64	–	–	–
All human-19	0.62	–	–	–
Rodent-19	0.55	–	–	–

Parenthetical numbers denote total number of siRNAs in given data sets. 'all' denotes data from the parental set of siRNAs, 'All human' denotes data from human sequences, 'Human E2s' denotes data from human E2 genes, 'Rodent' means rodent data only, 'Random all (*n*)' means data points from a randomly selected set of n human sequences.

in the inactive siRNAs at each of these three individual positions (7C, 11C and 19A). An excess of G at position 21 in the 3′-overhang of potent siRNAs was observed, and was confirmed by an excess of no-G in weak siRNAs. This event was matched by an excess of C in inactive sequences at a similar frequency and statistical significance. This finding has not been reported previously, possibly in part because siRNAs are often used with a dTdT 3′- terminus. As the overhangs in our sequences were DNA, it is not clear whether there is a preference for ribo-G or for deoxyribo-G. Some motifs have *P* values near the Hochberg threshold and may become insignificant when the data subset is enlarged or decreased[22], but they may be important as part of more complex motifs. In summary, mining the data with statistical methods identified previously known motifs in potent siRNAs—which validates the data set—and also, new over-represented motifs.

Data were divided randomly into training and testing sets of 2,182 and 249 sequences, respectively. An ANN was trained and tested on the full ('All') training and full ('All') testing set, respectively. A scatter plot of experimentally determined activity versus predicted activity produced a Pearson coefficient of correlation $r = 0.66$ (**Fig. 2c**: $P = 2.2 \times 10^{-16}$) and only a slightly higher correlation of 0.67 when applied to the training set itself (data not shown). This proves that computational intelligence can be used to predict with accuracy siRNA potency based upon nucleotide sequence alone. The fraction of siRNAs showing the highest predicted inhibition ($N = 62$; algorithm score of 0.75–0.85) also returned the highest experimental inhibition (mean = 84%; s.d. = 14.7%). Conversely, the fraction of siRNAs showing the lowest predicted activity ($N = 29$; algorithm score of 0.25–0.35) returned a mean inhibition in the reporter assay of 47% (s.d. = 14.9%). The two standard deviations are similar and indicate that the algorithm does not predict high activity more accurately than low activity. We calculated a number of specificities and sensitivities for the ANN, hereafter called BIOPREDsi (data from **Fig 2c**), with the view that maximum algorithm specificity (at the expense of sensitivity) is usually required during selection of siRNAs, so as to minimize the probability of false positives. Thus, for an algorithm score (Biopred score) of ≥ 0.75, the top 10% most potent siRNAs (from the test set of 249 siRNAs) yield a specificity of 79% and a sensitivity of 53%, whereas the top 33% provide a specificity of 83% and 43% sensitivity, respectively.

Subsets of the data were used to train and test additional ANNs to probe the limits of algorithm performance (**Table 1**). This might have revealed if bias had been introduced during selection of the 34 target sequences, for example, from use of a large number of E2 sequences, or from inclusion of rodent sequences. With a constant testing set ('All'), Pearson coefficients increased with training set size and reached a plateau at ~700 data points (**Fig. 2d**). Some predictive power was available with only 218 siRNAs, probably indicating the dominance of a few motifs. This was also observed with the other test sets. The best performing predictor on all test sets was obtained from training with the 'All' (2,182 siRNAs) group. Exclusion of rodent sequences from training and testing sets did not unduly influence algorithm performance, nor did a 'Human E2' predictor perform better with the 'h E2' set. The algorithms were particularly effective on the rodent test set, perhaps a consequence of the relatively small size of the set (51 siRNAs). Three ANNs were trained using nucleotides 1–19 of the 21-nucleotide guide strand (**Table 1**); performance of the two 'All-19' predictors was inferior to that of their 21-nucleotide counterparts on the 'All' testing set, whereas that of 'Rodent-19' was equal to its 21-nucleotide counterpart, suggesting that complementarity over the full oligoribonucleotide guide length provides an improved gene knock-down.

The algorithms' ability to rank siRNAs targeted to endogenously expressed mRNAs was assessed using the three previously used data sets (*TC10*, *UBE2I*, *CDC34*). For each siRNA, the Biopred score was plotted against experimentally determined activity, and the following correlation coefficients were returned: for *TC10* siRNAs, $r = 0.60$ ($N = 12$, $P = 4.0 \times 10^{-2}$; **Fig. 3a**) and for *UBE2I* and *CDC34* siRNAs, assayed at the protein level, $r = 0.60$ ($N = 14$, $P = 2.4 \times 10^{-2}$; **Fig. 3b**), and $r = 0.77$ ($N = 10$, $P = 9.9 \times 10^{-3}$; **Fig. 3c**; ($r = 0.36$, $P = 0.29$ with the outlier mentioned in the figure legend included), respectively. Although proper evaluation of BIOPREDsi performance requires a data set of appropriate size and unbiased content, in these three cases (total 36 siRNAs), good to excellent Pearson correlation coefficients were obtained ($r = 0.60$–0.77). The performances of several siRNA selection tools were recently cross-compared using three fairly large, published data sets generated by different methods, in which inhibitory activities are reported for 19-nt sequences, that is, 21-nt siRNAs most of which bear noncomplementary dTdT overhangs[11]. Algorithm All-19, (training set derived from 2,182 data points), returned Pearson correlations of 0.55, 0.57 and 0.45 for the data sets of Reynolds (240 siRNAs)[7], Vickers (76 siRNAs)[23] and Horbarth (44 siRNAs)[24], respectively, scores superior to those of the other reported algorithms. BIOPREDsi was then tested in its intended application: for six genes, the two most active predicted, specific siRNAs (see **Supplementary Table 5** online) were identified and screened using Q-PCR at three doses as single reagents, and also, as a mixture (**Fig. 3d–i**). In each case an effective inhibition was observed at the maximum dose and in general, the activity of the mixture was not significantly better than that of single reagents.

Finally, BIOPREDsi was used in the design of a genome-wide library of 48,746 siRNAs. A comprehensive description of library design will be published elsewhere. Briefly, 24,373 target genes (contigs) were selected from a proprietary database of predicted coding transcripts based on mRNA and EST evidence and, also, on the presence of splice sites. Transcript coverage for each contig was prioritized and maximized by tiling candidate siRNAs across all common exons of high-confidence transcripts. Where coverage of all transcripts was not possible (~8%), another group of candidate siRNAs was created, such that combination of the two candidate siRNA groups targeted the maximal set of high-confidence transcripts.

siRNA sequences were then filtered according to predicted specificity; for both guide and sense strands, a maximum number of mismatched nucleotides and a minimum length of nucleotide identity to all other likely alternative binding sites in the transcriptome were required. Remaining candidates were then ranked according to the Biopred score and the two predicted most potent siRNAs (score $\geq$0.75 for 96% of cases) were acquired for the library. Published reports[25] and our experiments suggest that the gene silencing potential of a mixture of two siRNAs is not inferior to that of the most potent of the individual siRNAs. Therefore, we pooled two siRNAs for the 24,373 targets to interrogate the pathway mediated by the hypoxia-inducible factor HIF-1A that allows mammalian cells to adapt to low oxygen levels. Under hypoxic conditions HIF-1α levels increase, the protein translocates to the nucleus and heterodimerizes with HIF-1β (ARNT). Coactivators CBP and p300 are recruited to the complex and increase the transcription of genes involved in glucose metabolism, angiogenesis and erythropoiesis. HIF-1A activity was measured by the use of a firefly luciferase reporter gene bearing three copies of the erythropoietin enhancer containing the core hypoxia-response element (HRE). The activity of this reporter is increased up to 50-fold after exposure to 1% oxygen or 100 μM desferioxamine mesylate, which mimics hypoxia by inactivating HIF-1A-modifying prolyl hydroxlases. The screen was done in duplicate in HeLa cells. siRNAs targeted to the GL3 luciferase reporter gene, present on each of the microtiter-plates, inhibited $\geq$70% of enzymatic activity in ~99% of these wells, as compared to the median of the noncontrol samples on each plate. siRNAs targeting the two essential mediators of the response to hypoxia, HIF1α and ARNT, scored among the top 24,373 siRNA

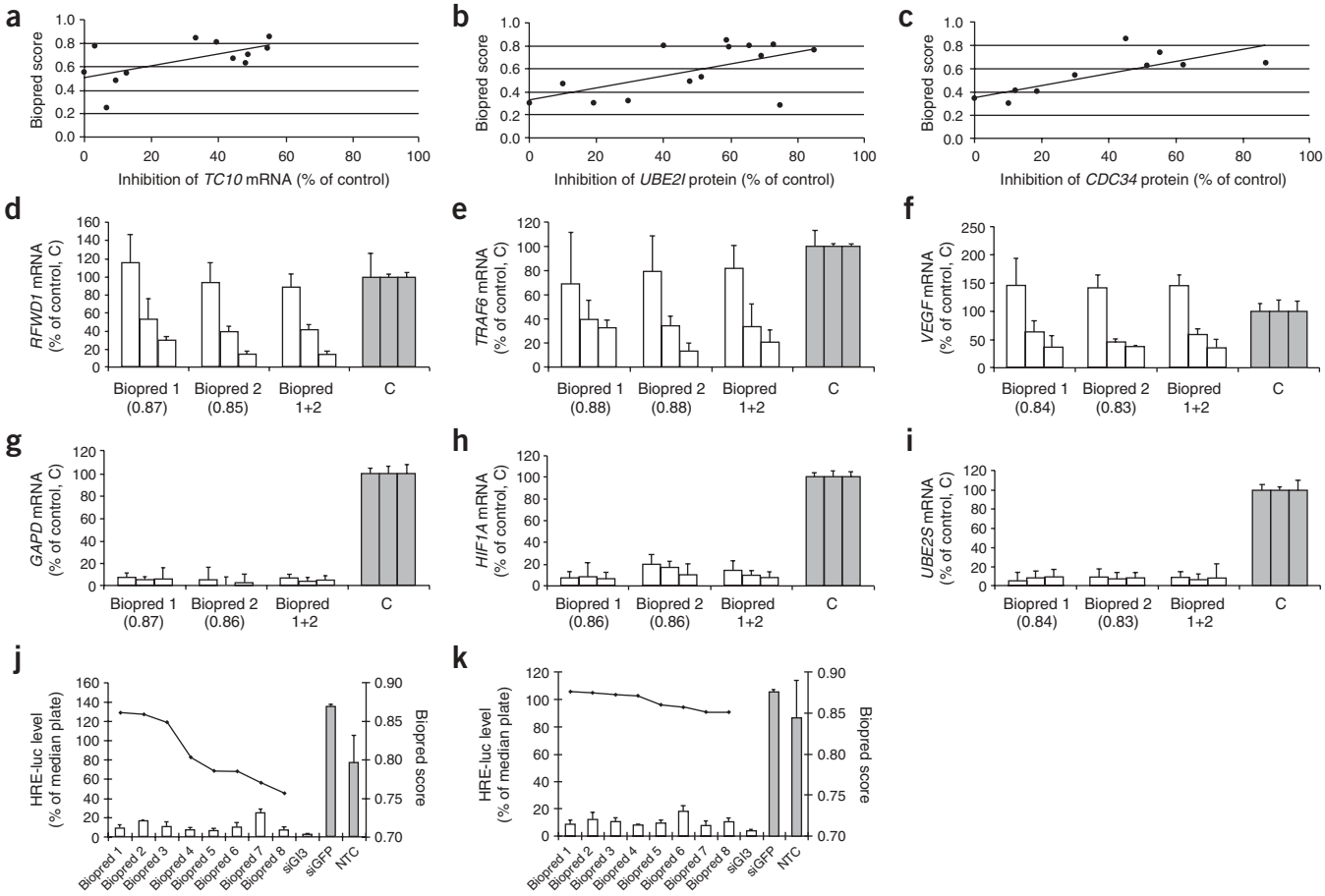

Figure 3 Performance of the algorithm with respect to both ranking 36 siRNAs of various potencies and selecting potent siRNAs. (**a**) Two-dimensional plot with linear regression line of predicted activity of twelve siRNAs targeted to *TC10* ras-like gene against observed normalized inhibition of mRNA (least potent siRNA set to 0% mRNA inhibition): Pearson coefficient: $r = 0.60$. (**b**) Two-dimensional plot with linear regression line of predicted activity of fourteen siRNAs targeted to *UBE2I* against observed inhibition of protein (western blot measurements were normalized to the weakest siRNA, arbitrarily set to 0% inhibition): Pearson correlation coefficient: $r = 0.60$. (**c**) Two-dimensional plot with linear regression line of predicted activity of ten siRNAs targeted to *CDC34* against observed inhibition of protein (western blot measurements were normalized to the weakest siRNA, arbitrarily set to 0% inhibition): Pearson correlation coefficient: $r = 0.77$. The algorithm was used to select the top two predicted siRNAs against six human genes. siRNAs were assayed for downregulation of target siRNA by Q-PCR separately and also as an equimolar mixture at three concentrations, and normalized to luciferase siRNA (see **Supplementary Tables 5–7** online). Individual algorithm scores are marked below the bars. (**d**) Inhibition of *RFWD1* (NM_032271) at 5, 15, 45 nM concentrations in HeLa cells. (**e**) Inhibition of *TRAF6* (NM_004620) at 5, 15, 45 nM concentrations in HeLa cells. (**f**) Inhibition of *VEGF* (NM_003376) at 20, 40, 60 nM concentrations in H1299 cells. (**g**) Inhibition of *GAPD* (NM_002046) at 5, 15, 45 nM concentrations in HeLa cells. (**h**) Inhibition of *HIF1A* (NM_001530) at 20, 40, 60 nM concentrations in H1299 cells. (**i**) Inhibition of *UBE2S* (NM_014501) at 20, 40, 60 nM concentrations in H1299 cells (**j**) Inhibition of HRE-luciferase activity by eight siRNAs targeted to *HIF1A* (NM_001530), in HeLa cells (including the two siRNA sequences employed in the primary screen). (**k**) Inhibition of HRE-luciferase activity by eight siRNAs targeted to *ARNT* (NM_001668) in HeLa cells (including the two siRNA sequences employed in the primary screen).

pools (ranking positions 6 and 168, respectively), reproducibly inhibiting HRE-Luc activity by ≥74% compared to plate median. For both genes, inhibition of HRE luciferase activity and expression of endogenous mRNAs and proteins was confirmed with each of the eight most potent predicted siRNAs (Biopred scores of 0.86–0.75 and 0.86–0.84 respectively; **Fig. 3j,k**). In summary, the validated results of the high-throughput reporter gene assay of siRNA function have provided a sound empirical basis for training ANNs and for the development of a powerful siRNA design system. This has proven its potency in guiding genome-wide screening operations. Application of the reporter assay facilitates identification of oligonucleotide properties that are only revealed through analysis of large data sets[26].

METHODS

siRNAs. siRNAs were provided by Qiagen AG as 21-nt oligoribonucleotides with a 19 base pair duplex region and two deoxynucleotide overhangs on the 3′-terminus of each strand. The DNA of the sense strand was a dTdT, whereas the overhang of the antisense strand was complementary to the target mRNA. An siRNA targeting the 5′-UTR region of eYFP mRNA (NAS-12842) was used as a positive control, and NAS-8549 was used as a common negative control. The sequences of all siRNAs are listed in **Supplementary Table 3**.

Cell culture. Human NCI-H1299 and HeLa cells obtained from ATCC were maintained in 5% humidified CO_2 atmosphere at 37 °C in RPMI1640 and DMEM medium (Life Technologies), respectively, supplemented with 10% (vol/vol) fetal bovine serum (FBS). Subconfluent cells were washed, trypsinized and plated into the assay plates in media without antibiotics 24 h before transfection.

Reporter expression clones. The eCFP-eYFP dual reporter vector pNAS-092 was described previously[20]. It contains a multiple cloning site after the stop codon of the eYFP for inserting the appropriate cDNAs of interest. pNAS-092 was converted to the GATEWAY destination vector pNAS-156 by inserting the attR1 and attR2 cloning sites after the YFP stop codon as recommended by the manufacturers (Invitrogen). Subsequently, cDNA target segments were inserted by ligation (pNAS-092) or recombination (pNAS-156) and verified by sequencing. The inserts were 344–3,784 nucleotides in length and were in most cases from coding regions; the remainder were UTR. The sequences of the cDNA inserts in the reporter plasmids are listed in **Supplementary Table 2**.

siRNA transfection and dual-reporter assay. H1299 cells were seeded in Costar 96-well assay plates and cotransfected with the reporter plasmid using Lipofectamine/Plus reagent according to the manufacturer's instructions (Invitrogen; final concentration: 1 ng/μl plasmid, 1.3 μl/ml Lipofectamine). After 2 h, siRNAs were added in the presence of Oligofectamine (Invitrogen; final concentration: 8 μl/ml Oligofectamine/50 nM siRNA). After a further 2-h incubation at 37 °C the medium was removed and replaced with 100 μl RPMI medium without phenol red supplemented with 10% (vol/vol) FBS and incubated for 3 d at 37 °C. The fluorescent read-out was measured in 24 h intervals over 3 d in a plate reader (Victor, Berthold Technologies) as described previously. Fluorescence of eCFP and eYFP was measured with the excitation filter of 436/20 nm and 500/25 nm and with the emission filter of 480/30 nm and 535/30 nm respectively. The quotient of eYFP/eCFP fluorescence counts represents inhibition of eYFP.

Western blot analysis. Western blot analysis was performed as described[27]. Protein quantification was conducted with the BCA method (PIERCE). Gel electrophoresis was performed on a 4–12% NuPAGE bis-Tris gel (Invitrogen) with 10 μg protein per slot. Proteins were electro-blotted to PVDF Immobilon-P membranes (Millipore) and developed using the mouse monoclonal antibodies against CDC-34 (ref. 28) and uBE21 (1:500; Pharmingen, #610748) in combination with anti-alpha-tubulin (1:10,000, Sigma, #T-5168), which served as a loading control. Horseradish peroxidase-coupled goat anti-mouse IgG secondary antibody (Sigma, #A2304) was used at a 1:1,000 dilution.

mRNA analysis with Q-PCR. Total RNA was prepared using the RNeasy 96 kit (Qiagen #74183) according to the manufacturer's instructions and quantified by OD_{260} measurement (Spectramax). Primer pairs and FAM-labeled TaqMan probes for real time PCR were designed using the Primer Express v1.0 program (ABI PRISM, PE Biosystems) and purchased from Microsynth (Switzerland). Alternatively, "Assays-on-demand" primers were purchased from Applied Biosystems. The primer sequences are listed in **Supplementary Table 6**. RNA samples were assayed either with the Reverse Transcriptase Q-PCR Mastermix kit (RT-QPRT-032X, Eurogentec) or, for "assay on-demand" kits, with the RT-PCR Mastermix (Applied Biosystems, #4309169) using an ABI PRISM 5700 (Applied Biosystems).

Screening and normalization of data for ANN training. YFP fluorescence for each siRNA was assayed on duplicate plates and at two time points. The whole plate was discarded if for any time point the duplicate plates did not show a Pearson coefficient of correlation of ≥0.7 or when the positive control siRNA did not inhibit YFP expression by ≥60%, compared to negative control. Data for 2,675 siRNA sequences passed this filter. Subsequently, 9% of the data was discarded, because duplicate measurements of two time points for individual siRNAs showed a standard deviation of ≥30% inhibition. The remaining duplicate data points were averaged and used for neural net training and testing. The final data set contained 2,431 sequences. For the training all data were normalized against the negative control (set at 10% activity) and the positive control (at 90% activity) which differs from the normalization used for reporter data plots and directly mentioned in the text. Correlation is invariant under these two normalization variants. For the network training, excess activity outside of a 0–100% range was dampened down to fit this activity range. The output of the network is given as a fraction between 0–1.

Generation of training and testing sets. The training/testing division was performed with [pseudo-] random numbers picked for each siRNA with a chance of 1:9 of assignment to the training set (2182 siRNA) or testing set (249 siRNA). Four differently sized subsets were constructed by randomly picking siRNAs from the largest training set. Additional subsets were constructed by separating the full size sets of training and testing by their target origin (that is, human, rodent, E2).

Method of training ANNs. Data from a given training set were submitted to network training using the Stuttgart Neural Net Simulator (SNNS) (http://www-ra.informatik.uni-tuebingen.de/SNNS/). The siRNA sequence was presented to the input layer and the reporter data were used to adjust the weights between the network nodes. Each siRNA sequence and its target inhibition value were presented a total of ten times. After a one time presentation of all data points the weights of the network were updated synchronously with a learning rate of 0.1. Based on five different initializations of weights, the resulting weights of five trained networks differed but all five networks consistently showed only slightly varying prediction output. A final output was achieved by averaging the signals of the respective output node of all five networks.

High-throughput siRNA transfection for luciferase reporter gene assay. Each of nine 150 cm^2 flasks of HeLa cells were transfected with 50 μg of HRE-Luc reporter plasmid (a gift of David M. Livingston, Dana-Farber Cancer Institute) and 10 μg of pRL-SV40 Renilla Luciferase plasmid (Promega), using 180 μl FuGENE6 transfection reagent (Roche), according to manufacturer's instructions. Plasmid-transfected cells were trypsinized, counted using a ViCell XR (Beckman Coulter) and diluted in medium to a concentration of 50,000 cells/ml; 25 μl of the cell suspension, containing 1,250 cells, was then dispensed to each well of an opaque, white, 384-well microtiter plate (Nalge Nunc) using a Multidrop (Thermo) and cultured for 24 h. For siRNA transfection, Oligofectamine (Invitrogen) was diluted to a concentration of 0.07 μl oligofectamine per 4 μl Opti-MEM I, and incubated for 5 min. Following the incubation, 4 μl of diluted transfection reagent was added to 4 μl of a 375 nM siRNA pool of two duplexes, diluted in Opti-MEM I, in a 384-well storage plate (Nalge Nunc). The 8 μl siRNA transfection mixture was incubated for 20 minutes and then transferred to a 384-well microtiter plate of HeLa cells. HeLa cells were treated with 100 mM Deferoxamine Mesylate (Sigma) 48 h after siRNA transfection. After 24 h, cells were assayed for Firefly and Renilla luciferase activity using the Dual-Glo Luciferase Assay System (Promega), per manufacturer

instructions, and an EnVision Reader, with 100 millisecond integration per well (Perkin Elmer). BIOPREDsi is available at http://www.biopredsi.org/.

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

ACKNOWLEDGMENTS

We thank Erik Bury for preliminary work on development of the algorithm. We also thank B. Csordas, V. Drephal, A. Garnier, S. Gfeller, D. Kirk, B. Pak, R. Theurillat, R. Widmer, S. Zhao and W. Zuercher for excellent technical support. We thank J. Mestan and F. Hofmann for E2 antibodies. We thank P. Weiss for helpful discussions and J. Hunziker and R. Haener for carefully reading the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Received 10 January; accepted 27 April 2005
Published online at http://www.nature.com/naturebiotechnology/

1. Paddison, P. *et al.* A resource for large-scale RNA-interference-based screens in mammals. *Nature* **428**, 427–431 (2004).
2. Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431–437 (2004).
3. Kittler, R. *et al.* An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. *Nature* **432**, 1036–1040 (2004).
4. Boese, B. *et al.* Mechanistic insights aid computational short interfering RNA design. *Methods Enzymol.* **392**, 73–96 (2005).
5. Khvorova, A., Reynolds, A. & Jayasena, S.D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216 (2003).
6. Schwarz, D.S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208 (2003).
7. Reynolds, A. *et al.* Rational siRNA design for RNA interference. *Nat. Biotechnol.* **22**, 326–330 (2004).
8. Hsieh, A.C. *et al.* 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 (2004).
9. Ui-Tei, K. *et al.* Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.* **32**, 936–948 (2004).
10. Amarzguioui, M. & Prydz, H. An algorithm for selection of functional siRNA sequences. *Biochem. Biophys. Res. Commun.* **316**, 1050–1058 (2004).
11. Saetrom, P. & Snove, O., Jr. A comparison of siRNA efficacy predictors. *Biochem. Biophys. Res. Commun.* **321**, 247–253 (2004).
12. Labuda, D., Nicoghosian, K. & Cedergren, R.J. A novel RNA digesting activity from commercial polynucleotide phosphorylase. *FEBS Lett.* **179**, 213–216 (1985).
13. Kierzek, R. Hydrolysis of oligoribonucleotides: influence of sequence and length. *Nucleic Acids Res.* **20**, 5073–5077 (1992).
14. Schneider, G. & Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **70**, 175–222 (1998).
15. Weinstein, J.N. *et al.* Neural computing in cancer drug development: predicting mechanism of action. *Science* **258**, 447–451 (1992).
16. Stolorz, P., Lapedes, A. & Xia, Y. Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.* **225**, 363–377 (1992).
17. Giddings, M.C. *et al.* Artificial neural network prediction of antisense oligodeoxynucleotide activity. *Nucleic Acids Res.* **30**, 4295–4304 (2002).
18. Chalk, A.M. & Sonnhammer, E.L. Computational antisense oligo prediction with a neural network model. *Bioinformatics* **18**, 1567–1575 (2002).
19. Rumelhart, D. in *Parallel distributed processing* (eds. McClelland, J. & the PDP Research Group), vol. 1, 318–362, (MIT Press, Cambridge, MA).
20. Huesken, D. *et al.* mRNA fusion constructs serve in a general cell-based assay to profile oligonucleotide activity. *Nucleic Acids Res.* **31**, e102/1–e102/11 (2003).
21. Zamore, P.D., Sharp, P.A., Tuschl, T. & Bartel, D.P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
22. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. of the Royal Statistical Society. Series B* **57**, 289–300 (1995).
23. Vickers, T.A. *et al.* Efficient reduction of target RNAs by small interfering RNA and RNASE H-dependent antisense agents. *J. Biol. Chem.* **278**, 7108–7118 (2003).
24. Horbarth, J. *et al.* Sequence, chemical and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. *Antisense Nucleic Acid Drug Dev.* **13**, 83–106 (2003).
25. Ohba, H. *et al.* Inhibition of bcr-abl and/or c-abl gene expression by small interfering, double-stranded RNAs: cross-talk with cell proliferation factors and other oncogenes. *Cancer* **101**, 1390–1403 (2004).
26. Hall, J. Unraveling the general properties of siRNAs: strength in numbers and lessons from the past. *Nat. Rev. Genet.* **5**, 552–557 (2004).
27. Hemmings-Mieszczak, M., Dorn, G., Natt, F., Hall, J. & Wishart, W. Antisense oligonucleotides complement RNAi-mediated specific inhibition of the recombinant rat P2X3 receptor. *Nucleic Acids Res.* **31**, 2117–2126 (2003).
28. Butz, N. *et al.* The human ubiquitin-conjugating enzyme Cdc34 controls cellular proliferation through regulation of p27Kip1 protein levels. *Exp. Cell Res.* **303**, 482–493 (2005).