Published online 14 December 2012 *Nucleic Acids Research, 2013, Vol. 41, No. 3* ***1383–1394***
doi:10.1093/nar/gks1191

Improved nucleic acid descriptors for siRNA efficacy prediction

Simone Sciabola[1,*], Qing Cao[1], Modesto Orozco[2,3,4], Ignacio Faustino[2,3,4] and Robert V. Stanton[1]

[1]Pfizer Oligonucleotide Therapeutic Unit, 620 Memorial Drive, Cambridge, Massachusetts 02139, United States, [2]Institute for Research in Biomedicine (IRB Barcelona), 08028 Barcelona, Spain, [3]Joint IRB-BSC Program in Computational Biology, 08028 Barcelona, Spain and [4]Department of Biochemistry and Molecular Biology, University of Barcelona, 08028 Barcelona, Spain

Received July 20, 2012; Revised October 9, 2012; Accepted October 29, 2012

ABSTRACT

Although considerable progress has been made recently in understanding how gene silencing is mediated by the RNAi pathway, the rational design of effective sequences is still a challenging task. In this article, we demonstrate that including three-dimensional descriptors improved the discrimination between active and inactive small interfering RNAs (siRNAs) in a statistical model. Five descriptor types were used: (i) nucleotide position along the siRNA sequence, (ii) nucleotide composition in terms of presence/absence of specific combinations of di- and trinucleotides, (iii) nucleotide interactions by means of a modified auto- and cross-covariance function, (iv) nucleotide thermodynamic stability derived by the nearest neighbor model representation and (v) nucleic acid structure flexibility. The duplex flexibility descriptors are derived from extended molecular dynamics simulations, which are able to describe the sequence-dependent elastic properties of RNA duplexes, even for non-standard oligonucleotides. The matrix of descriptors was analysed using three statistical packages in R (partial least squares, random forest, and support vector machine), and the most predictive model was implemented in a modeling tool we have made publicly available through SourceForge. Our implementation of new RNA descriptors coupled with appropriate statistical algorithms resulted in improved model performance for the selection of siRNA candidates when compared with publicly available siRNA prediction tools and previously published test sets. Additional validation studies based on in-house RNA interference projects confirmed the robustness of the scoring procedure in prospective studies.

INTRODUCTION

RNA interference (RNAi) has become an essential tool in functional genomics by enabling genome-scale loss of function screens for the identification of new drug targets (1). However, development of RNAi as a viable therapeutic has progressed more slowly, as the potential advantages over small molecules are balanced with the challenges of intracellular tissue specific delivery. Compounds that have advanced to the clinic have shown promise in the control of a wide array of disorders, including cancer, infectious diseases, cardiovascular, neurodegenerative, and obesity (2).

RNAi is an evolutionary conserved, efficient and specific pathway by which short double-stranded RNAs (dsRNAs) trigger the inhibition of gene expression post-transcriptionally (3,4). RNAi can be endogenously processed and expressed or exogenously introduced with chemically synthesized small interfering RNA (siRNA). When in the cytoplasm, longer dsRNAs are processed by Dicer, an RNAse III endonuclease enzyme, which cleaves into 21- or 22-nucleotide-long dsRNA molecules, with 3′-overhangs (nucleotides that do not form part of the duplex) of 2 nucleotides on the end of each the sense and antisense strands. Current models hypothesize that Dicer selects cleavage sites by measuring a set distance from the 3′-overhang of the dsRNA terminus (5). Dicer is also required for loading and function of the RNA-induced silencing complex (RISC). Recent reports suggest that RISC must be activated from a latent form, containing a double-stranded siRNA, to an active form, RISC*, through unwinding of the siRNAs. This is controlled by one of the RISC's functional units, Argonaute 2 (Ago2), which recognizes and cleaves the passenger strand of the

*To whom correspondence should be addressed. Tel: +1 617 551 3327; Fax: +1 845 474 3572; Email: simone.sciabola@pfizer.com

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.

siRNA, hence releasing the guide strand from the duplex (6–8). The activated RISC then uses the unwound siRNA as a guide to trigger sequence-specific mRNA degradation. Although theoretically any 21-nucleotide region of an mRNA can be used as the basis of design for an siRNA, in practice, different intrinsic activities are seen. This difference in activity can be attributed to many reasons, including hybridization energy, mRNA secondary structure motifs and any other components coded in the sequence. Algorithms for the selection of efficient and selective siRNA sequences are therefore necessary, and a number of tools have been developed for predicting siRNA efficacy, with varying accuracy. These predictions tools can be classified into two different groups: (i) rule-based and (ii) machine-learning.

First-generation siRNA design tools were developed through the study of small data sets and consist of guidelines in contrast to a quantitative scoring scheme. Tuschl *et al.* (9) were among the first to come up with a set of siRNA design rules, based on G/C content and symmetric 3′ TT overhangs. Khvorova *et al.* (10) found that functional siRNA duplexes displayed a lower internal duplex stability at the 5′-end of the antisense strand when compared with non-functional duplexes, suggesting the key role played by duplex thermodynamics in biasing strand selection during siRNA–RISC assembly and activation. Amarzguioui and Prydz (11) confirmed the importance of duplex end stability asymmetry and identified sequence motifs on the siRNA sense strand that consistently correlated positively (G-/C-1, A-6, A-/U-19, where the number corresponds to the position on the antisense strand) or negatively (U-1, G-19) with functionality across a data set of 80 siRNAs targeting four genes. With a data set of 62 siRNAs targeting four exogenous and two endogenous genes, Ui-Tei *et al.* (12) derived four design rules:(i) the use of an A/U at the 5′ end of the antisense strand, (ii) G/C at the 5′ end of the sense strand, (iii) enrichment of A/U residues in the terminal one-third of the antisense strand and (iv) absence of any GC stretch of >9 nucleotides in length. Reynolds *et al.* (13) published a set of eight rules based on a systematic analysis of 180 siRNAs that stressed the importance of duplex thermodynamics, as determined by overall GC content, lower stability at 3′-end of the sense strand, controlling for potential internal hairpins and presence of a uridine at position 10 of the sense strand.

Second-generation machine learning-based algorithms were first introduced by Huesken *et al.* (14). In their work, Huesken published a data set of 2431 randomly selected siRNAs targeting 34 mRNA species, which were consistently assayed through a fluorescent reporter gene system. This data set was subsequently used as the basis of siRNA efficacy models such as BIOPREDsi (14), DSIR (15), Thermocomposition (16) and i-Score (17). These machine learning models use a numerical description of the siRNA sequence features, which are statistically analysed with regression or classification algorithms. Whereas statistical algorithms vary in performance, the accuracy of siRNA models is primarily dependent on the descriptors and their degree of abstraction from the sequence information. Current tools are typically based on positional features, nucleotide composition, thermodynamics, energy profiles and local target mRNA stability. These descriptors primarily encode the nucleotide sequence of the siRNA [one-dimensional (1D) information] and in some cases the predicted mRNA secondary structure (two-dimensional information) (18–20). To date, three-dimensional (3D) structural information for the siRNA duplex has not been included as a descriptor in siRNA activity models. Additionally, RNA strain and flexibility have not been included in efficacy modeling studies, even though they have been shown to play a key role during binding of the siRNA guide strand to the Argonaute (Ago) silencing complex (21).

In this study, the use of sequence-based one-dimensional (1D) and structural 3D descriptors is investigated in siRNA efficacy models. The 1D descriptors used include positional and composition features, together with duplex thermodynamics and the auto-cross-covariance (ACC) transform description of the siRNA guide strand (22). The physical 3D structural descriptors include global helical stiffness, sequence-adapted stiffness parameters for the RNA duplex, which were derived from the analysis of molecular dynamics (MD) simulations on a set of diverse RNA duplexes (23). The information content and biological relevance of the 3D descriptors is studied through comparison with previous results. Using these descriptors and the Huesken data set (14), siRNA efficacy models were generated using three regression techniques: (i) partial least squares (PLS), (ii) random forest (RF) and (iii) support vector machine (SVM). Validation of the results was done through cross-validation and external predictions based on independent test sets, allowing the identification of the best combinations of descriptors and regression algorithm. The final model is available as part of the PFizer Rnai Enumeration and Design (PFRED) (H. Xi *et al.*, unpublished result) OpenSource project developed for the design, analysis and visualization of antisense and siRNA oligonucleotides.

MATERIALS AND METHODS

Data sources

Public data

Several sources of siRNA data were used in this study. The Huesken *et al.* (14) data set consisting of 2431 randomly selected siRNAs targeting 34 different mRNA transcripts was used to train and validate the model (14,24). Three additional data sets provided by Reynolds *et al.* (13) (244 siRNAs targeting seven genes), Vickers *et al.* (25) (76 siRNAs targeting the mRNA transcripts of CD54 and PTEN) and Harborth *et al.* (26) (44 siRNAs targeting Lamin A/C mRNA) were used to benchmark the model against other prediction algorithms.

In-house data

The prediction model developed here was used to design a set of 352 siRNA and 591 Dicer substrates. The siRNA were designed as 21-nucleotide dsRNA molecules, with 3′-overhangs of 2 nucleotides at the end of the each sense and antisense strand, whereas the Dicer substrates

were in either the R-Dicer or L-Dicer pattern. The R-Dicer consisted of 25 nucleotides in the sense strand and 27 in the antisense, with two DNA nucleotides at the 3′ end of the sense strand and a two-nucleotide overhang at the 3′-end of the antisense strand, and L-Dicer had 27 nucleotides in the sense strand and 25 in the antisense, with two DNA nucleotides at the 3′-end of the antisense strand and a two-nucleotide overhang at the 3′-end of the sense strand. The design patterns for the siRNA as well as R- and L-Dicer substrates are shown in Figure 4. All siRNA duplexes were either purchased from Integrated DNA Technologies (IDT) or synthesized in-house using standard protocols on a MerMade-192 synthesizer with 2′-TBDMS phosphoramidites and fast-base deprotecting group protocols at 200-nmole scale on CPG supports. Monomers were obtained from ChemGenes Corporation or Glen Research. After synthesis, the DMTr-off oligoribonucleotides were cleaved from the support and deprotected using AMA (a 50:50 mixture of ammonium hydroxide and aqueous methylamine) at 65°C for 1 h. The base-deprotected oligoribonucleotides were desilylated using TEA-HF/NMP per the Wincott *et al*. procedure (27). In most cases, the crude desalted oligoribonucleotides were of sufficient purity. If necessary, further purification was done by reverse-phase high-performance liquid chromatography and desalting with cartridge-based methods. The final oligoribonucleotides were characterized using a Waters Acquity ultra-performance liquid chromatography connected in-line to a Waters LCT Premier ToF mass spectrometer. The final oligoribonucleotides were dissolved in IDT duplex buffer (100 mM potassium acetate, 30 mM HEPES, pH 7.5) to yield a stock solution at 60 μM. Equal volumes of the complimentary oligoribonucleotides at 60 μM were mixed together and the resulting solution was heated to 90–100°C for 5–10 min then slowly cooled to room temperature.

Data were generated for hepatocellular carcinoma (HCC) relevant genes: hypoxia-inducible factor 1 alpha subunit (HIF1A), hexokinase-2 (HK2), heparanse (HPSE), survivin (BIRC5), histone-lysine N-methyltransferase (EZH2), c-Myc (MYC), FK506 binding protein 12-rapamycin associated protein 1 (MTOR), beta-catenin (CTNNB1), proto-oncogene B-Raf (BRAF) and phosphoinositide-3-kinase catalytic alpha polypeptide (PIK3CA). Hep3B cells (American Type Culture Collection) were grown in EMEM (ATCC) supplemented with 10% fetal calf serum. Cells were maintained in monolayer cultures at 37°C in an incubator with 5% CO_2. One day before transfection, cells were seeded at 5000 cells per well in 96-well plates. Cells were transfected at 30–60% confluence using Lipofectamine™ RNAiMAX (Life Technologies) according to the manufacturer's instructions and the indicated doses of each siRNA. The 21-mers siRNAs and Dicer substrates targeting HK2 and HIF1A were transfected at 0.08, 0.4, 2 and 10 nM concentrations. The compounds designed for HPSE were transfected at 0.08, 0.4 and 2 nM, as the hit rate was high enough at 2 nM to not require a higher concentration. These initial results were used to define a screening funnel for target validation, and only 5–10 *in silico* predicted siRNA screened at 10 nM would be necessary to identify a potent tool. These guidelines were used for all the follow-up studies on HCC relevant genes BIRC5, EZH2, MYC, MTOR, CTNNB1, BRAF and PIK3CA. Given that transfection efficiencies can vary according to the ratio of nucleotide to transfection reagent, each dose was supplemented with non-targeting (negative) control siRNA, such that the total RNA concentration was equal across experiments. RNA was purified 48 h post-transfection using the RNeasy mini kit from Qiagen. QuantiGene 2.0 assay (Affymetrix Inc. Santa Clara, CA) was used to measure the expression level of target genes before and after knockdown in Hep3B cell lines. Branched DNA probes for targeting genes and housekeeping gene PPIB probes were purchased from Affymetrix. Standard assay procedures were carried out according to the manufacturer's recommendations. Assay plates were read on the GloRunner Microplate Luminometer (Promega Corp, Sunnyvale, CA). The data reported in this study are normalized against the housekeeping gene PPIB. Cytotoxicity was not observed at the concentrations tested.

Molecular descriptors

By convention, the numerical descriptors of the siRNA sequence refer to the nucleotides in the guide (antisense) strand, which are ordered in the 5′ → 3′ direction from 1 to 21 (including the two-nucleotide overhang at position 20 and 21).

Sequence position

It can be shown that to characterize four different objects in an unbiased way (no particular object should be represented as being more similar or dissimilar with respect to the others), it is sufficient to use three indicator variables. The tetrahedron is a geometric representation that fulfills this condition, with the corner coordinates of the tetrahedron being used as qualitative sequence descriptors (Figure 1). In this way, the four nucleotides A, C, G and U (or T) can be placed in four selected, diametrically opposed, corners of a cube so that all the inter-objects distances are identical. The corresponding numerical representation for the four nucleotides will be as follows: A (−1, −1, +1), C (+1, −1, −1), G (−1, +1, −1) and U/T (+1, +1, +1). Each siRNA guide strand is then described by a numerical vector of length 63 (21 position in the guide strands × 3 indicator variables), which will be used to capture the frequency of the four nucleotides at specific positions along the sequence.

Sequence composition

The characterization of the siRNA guide strand in terms of global content of specific short nucleotide motifs was done using a count of occurrences of each nucleotide motif of length 1–3. Given that there are, respectively, 4, 16 and 64 potential nucleotide motifs with length 1, 2 and 3, the overall nucleotide content for each siRNA sequence could be encoded in an ordered vector of length 84 (4 + 16 + 64).

ACC transformation

To take into account the potential lack of independence between subsequent nucleotide positions along the siRNA

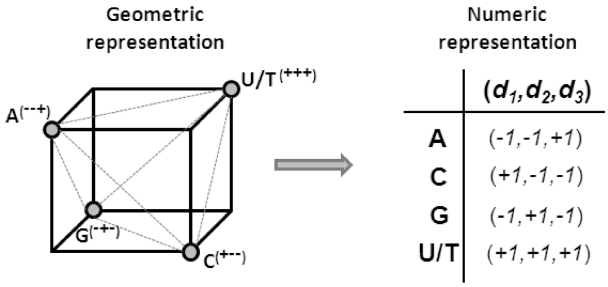

Figure 1. A perfect tetrahedron can be drawn by joining together the four diametrically opposed vertices of a cube. Each of the vertices can be assigned to a set of numerical coordinates if the tetrahedron is placed in a three-dimensional coordinate system, with the origin in the center of the cube.

guide strand, a numerical transformation based on neighbor effects was included. Auto- and cross-covariance functions were originally described and validated for modeling of peptide and DNA sequences where the difference in length limited the application of position-based descriptors (22). For a given lag 'd', the corresponding auto-covariance function of a descriptor 'x' over a sequence of length 'L' is given by:

$$A_{xx}(d) = \frac{\sum_{i=1}^{L-d} (x_i)(x_{i+d})}{L-d}$$

in the same way we can define the lagged cross-covariance function between two descriptors 'x' and 'y' as:

$$C_{xy}(d) = \frac{\sum_{i=1}^{L-d} (x_i)(y_{i+d})}{L-d}$$

where 'i' is the nucleotide sequence position index running from 1 to L, 'x' and 'y' are taken from the numerical representation previously described in 'Sequence Position' and 'd' (lag) is the maximum distance at which two nucleotides are allowed to interact (d = 13) in our study, giving an overall descriptor vector of length 108.

Duplex thermodynamics stability

Specific thermodynamic profiles of duplex RNA calculated at the sequence level have been shown to correlate well with siRNA functionality. Schwarz *et al.* (28) showed that the two strands of an siRNA duplex are not equally eligible for RISC assembly, and this asymmetry is a feature shared among both siRNAs and microRNAs (miRNAs). Based on a statistical analysis of the internal duplex stability of published miRNA precursors and siRNA sequences, Khvorova *et al.* (10) found that the 5′-AS region of the duplex siRNA was on average less stable than the 5′-S terminus in functional siRNAs. To take the thermodynamics effect into account, the INN-HB nearest neighbor (NN) model from Xia *et al.* (29) was used. The overall enthalpy (ΔH), entropy (ΔS) and free energy (ΔG) change on binding of the two siRNA strands were calculated, as well as the corresponding melting temperature (Tm). This resulted in a set of 11 differential end stability descriptors being added to the list to consider the free-energy differences (ΔΔG) between all the permutations of the first and last three base pair stacks ($\Delta\Delta G^{NN18\text{-}NN1}$, $\Delta\Delta G^{NN18\text{-}NN2}$, $\Delta\Delta G^{NN18\text{-}NN3}$, $\Delta\Delta G^{NN17\text{-}NN1}$, $\Delta\Delta G^{NN17\text{-}NN2}$, $\Delta\Delta G^{NN17\text{-}NN3}$, $\Delta\Delta G^{NN16\text{-}NN1}$, $\Delta\Delta G^{NN16\text{-}NN2}$, $\Delta\Delta G^{NN16\text{-}NN3}$, $\Delta\Delta G^{(NN17+NN18)\text{-}(NN1+NN2)}$, $\Delta\Delta G^{(NN16+NN17+NN18)\text{-}(NN1+NN2+NN3)}$). In the notation used here, NN1 refers to the first nearest neighbor pair at the 5′-end of the antisense strand, whereas the NN18 refers to the last nearest neighbor interaction pair at the 3′-end of the antisense strand. The average internal stability at the cleavage site (AIS) is the average of internal stability values for positions 9–14 on the antisense strand. Two additional duplex descriptors are included to quantify the differential stability between the 5′- and 3′-ends with respect to the centered positions in the siRNA duplex ($\Delta\Delta G^{NN18\text{-}NN10}$, $\Delta\Delta G^{NN1\text{-}NN13}$). The free-energy profile for the sense-antisense duplex resulted in 18 additional descriptors corresponding to the dinucleotide nearest neighbors present in the 19-nucleotide stem-loop. All together, the block of thermodynamics descriptors is encoded in a ordered vector of length 36.

Duplex flexibility

Including duplex flexibility into the descriptor set is hindered by the scarcity of experimental data from RNA duplex structures. As a surrogate, the results of MD simulations were used to generate flexibility descriptors. A recent study (23) of the RNA duplex flexibility has shown the parm99 force field (30) with the parmbsc0 (31) modification reliably reproduces the limited structural data available on dsRNAs. In that study, four different 18-mer duplex-RNA sequences were simulated containing many copies of the 10 unique base steps and making possible a reliable analysis of the sequence dependence of duplex flexibility. After a previously used protocol (32,33), all simulations were performed in the isothermal–isobaric ensemble (T = 298 K, P = 1 atm) for 150 ns to capture the near-equilibrium dynamic properties of the duplexes. The extensive MD simulations performed in that study provided the estimation of local (dinucleotide step) and global descriptors of RNA structure and flexibility. Average base pair step helical parameters were calculated using *Curves*+ (34) for the three translational [shift (f), slide (l), rise (s)] and the three rotational [tilt (t), roll (r), twist (w)] movements, while the associated stiffness matrix (Ξ) was derived by the inversion of the covariance matrix (⊂) obtained from the equilibrated part of the trajectory.

$$\Xi = E(\Delta X)^{-2} = k_B T \subset^{-1} = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{wr} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{wt} & k_{rt} & k_t & k_{st} & k_{tl} & k_{tf} \\ k_{ws} & k_{rs} & k_{st} & k_s & k_{ls} & k_{lf} \\ k_{wl} & k_{rl} & k_{tl} & k_{ls} & k_l & k_{lf} \\ k_{wf} & k_{rf} & k_{tf} & k_{lf} & k_{lf} & k_f \end{pmatrix}$$

where E is the energy associated with the deformation ΔX, k_BT is the Boltzmann temperature factor and k stands for the different stiffness constants defining the 36 elements of

the stiffness matrix. Thus, local helical deformations and their associated force constants are defined for each of the 10 representative dinucleotide steps. Associated flexibility values for the rotational helical parameters were determined as the inverse of the corresponding force constants, so the higher the force constant, the stiffer the corresponding deformation will be. Average local helical parameters roll (r) and tilt (t) for each base pair step were used to calculate the angle of axis deflection (θ) and its directionality (Φ) measured from the direction of the major groove (35).

$$\theta = (r^2+t^2)^{1/2}$$
$$\phi = \tan^{-1}(t/r) \qquad \text{for } (r > 0)$$
$$\phi = 180+\tan^{-1}(t/r) \qquad \text{for } (r < 0)$$

Average interaction energies (stacking and hydrogen bonding) were also derived from energy analysis of the snapshots collected during the MD simulations and were introduced in the model. Analysis of global deformations including global bending, tilt and roll were calculated using Madbend (36) from the corresponding local parameters associated to every dinucleotide step. An ordered vector of length 330 was therefore used to describe the structural flexibility of all the siRNA sequences in the data set.

mRNA secondary structure

Studies have shown that accessibility of local mRNA structures is an important determinant of the ability of a target region to promote efficient gene silencing (37–41). Although incorporating mRNA secondary structure into the descriptor set may improve model predictions, it has not been included in this work. This decision was made because only minimal improvement to model performance was seen in previous studies (19,42), and in-house validation studies of target secondary structure descriptors did not result in improvements over siRNA duplex descriptors alone (Supplementary Data, Table S3). Prediction of mRNA secondary structure is an inherently difficult problem and uncertainties related to currently available methods likely reduce their utility. As future algorithms improve mRNA structure predictions, the incorporation of these descriptors into efficacy models is likely to significantly improve their performance.

Statistical modeling

In this study, we decided to use regression instead of classification models, as they provide more information, additional flexibility and ease of evaluation. Using a continuous variable for siRNA efficacy also avoids the step of defining an arbitrary threshold for assigning compounds as active or inactive. Given the heterogeneity of our sequence descriptors and to keep variables with large variance from overshadowing other variables with small numerical scale, the descriptor matrix was scaled before running the regression algorithms. This was done using the scale = TRUE option in the R environment, which divides the centered columns of the descriptor matrix by their standard deviations.

Partial least squares

The partial least squares regression (PLSR) technique has been found to be useful when the number of independent variables is comparable with or greater than the number of data points. It is normally a powerful alternative for cases where the solution of the classical least squares problem does not exist or is unstable. This occurs in data sets with highly correlated descriptors (43). PLSR's objective is to summarize the variation in a data matrix in terms of a few essential and informative scores, also known as latent variables (LVs), which are mutually independent (orthogonal) and represent linear combinations of the original descriptors. Because LVs are chosen in such a way as to provide maximum correlation with the dependent variable, PLS models use the smallest possible number of descriptors to explain the underlying variance in the data set, providing enhanced model precision and stability. Data were imported in R and scaled as described previously. The PLSR package in R (44) was used to compute the top 10 LVs and the cross-validated R^2 applied to select the optimal number of PLS factors to include in the final model.

Support vector machines

The theory of SVM has been extensively described (45,46). SVM has been successfully applied to a number of data classification problems where the geometrical interpretation consists of choosing the N-dimensional hyperplane that optimally separates clusters of vectors in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other. A vector in an SVM represents a set of features that describes one case, and the vectors near the hyperplane are called the support vectors. The optimal separating hyperplane has a number of attractive statistical properties, which are detailed by Vapnik (46). The SVM method can also be used in regression, maintaining all the main features that characterize the maximal margin algorithm: a non-linear function is learned by a linear learning machine in a kernel-induced feature space while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space. Some advantages of SVM in comparison with other methods include: (i) a global and perhaps unique solution, (ii) a general solution and thus avoiding overtraining, (iii) a sparse solution and (iv) non-linear relations can be modeled. In our study, we used the SVM's interface to libsvm (47) as implemented in the R package e1071. Data were scaled within R and the epsilon-type regression machine was used in combination with the radial basis function kernel. A grid search over the two tunable parameter gamma (γ) and cost (c) was performed, and the best value selected based on 10-fold cross-validation. The selected parameters for γ and c were used to build the final model.

Random forest

RF models produce accurate predictions that should not overfit the data. The RF algorithm uses an ensemble of unpruned decision trees, each of which is built on a bootstrap sample of the training set using a randomly

selected subset of variables (48). Whereas standard trees are built by splitting each node using the best split among all variables, in RF, each node is split using the best among a subset of predictors randomly chosen at that node. A large number of trees are grown to maximum size without pruning, and aggregation is done to produce the final prediction. In the case of classification, this is done by assigning the object to the class predicted by the majority of the trees, whereas for regression problems, the average of the individual tree predictions is taken. In our study, we used the randomForest interface in R to the Fortran programs by Breiman and Cutler (http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm). Data were scaled within R and the n_{tree} parameter was set to 1000. For the m_{try} parameter, it has been observed that the performance of RF varies little over a wide range of values, except near the extremes, $m_{try} = 1$ or p, where p is the total number of variables. Therefore, we decided to use the default value, which will sample p/3 randomly selected variables as candidates at each split.

RESULTS AND DISCUSSION

A preliminary analysis was carried out to test the relevance of different groups of sequence descriptors (positional, composition, ACC transform, thermodynamics and 3D duplex flexibility), when used in combination with various machine learning algorithms (SVM, PLS and RF). The training set from Huesken *et al.* (14) consisting of 2182 siRNA sequences was used to build all the model combinations. The test set of 249 sequences randomly defined in the same work from Huesken was used to validate each specific combination. Table 1 shows the results obtained for the 18 models with the agreement between *in silico* predictions and experiments represented by the Pearson correlation coefficient. Among the three statistical algorithms tested, SVM gave slightly better correlation with experiment. As previously shown from studies on the same data set, position-dependent descriptors ($R_{svm} = 0.66$) and motifs composition ($R_{svm} = 0.60$) were found to correlate best with siRNA efficacy. This can be explained by the fact that certain regions in the duplex are involved in specific recognition events, which might only be encoded in specific nucleotide combinations. Thermodynamic features ($R_{svm} = 0.53$) and 3D duplex flexibility descriptors ($R_{svm} = 0.53$) were also found to be important in explaining the overall variance in the training set. As might be expected, a model generated using all sequence descriptors showed the best correlation with experiments ($R_{SVM} = 0.71$), indicating that none of the descriptors blocks were perfectly correlated and each provides some element of unique information predictive of efficacy. The results were independent of the specific statistical algorithm applied.

Next the relative importance of each specific sequence descriptor was determined using the training set of 2182 siRNAs from Huesken. A variable selection algorithm was applied to test whether a given variable should be included in the final model. The procedure consists of an iterative evaluation of the effects of individual variables on the model predictivity based on the validation of a number of reduced models. These reduced models are created using variables combinations selected according to a fractional factorial design (49). Table 1 shows the final number of descriptors for each block that survived the variable selection procedure (within brackets), as well as the recalculated correlation coefficients for all the model combinations (descriptor blocks and learning algorithms) using the reduced number of descriptors. Once selection was complete, 148 variables were chosen from the original set of 642 sequence descriptors. The name of the selected variables can be found in the Supplementary Table S2. Overall, the optimized models (SVM, PLS and RF using 148 variables) showed better correlation with experiment than those using the full set of descriptors, with SVM giving the best results (Table 1 in parenthesis and Figure 2). An attempt was made to investigate the performance of our SVM model under the same descriptor selection and optimization scheme presented previously when the 3D structural descriptors were not included. The SVM model built excluding the 3D descriptors gave a correlation between experimental and predicted siRNA activity equal to $R = 0.70$ (313 variables). After applying descriptor selection, the best model used 219 variables with a correlation coefficient of $R = 0.75$ (Supplementary Table S3). In comparison, the correlation coefficient for the optimized SVM model trained with all the descriptors, 3D included, was $R = 0.80$ (Table 1), demonstrating that additional information is present in the 3D duplex flexibility descriptors, which is not encoded by any of the other descriptor blocks based only on the primary sequence information.

Table 1. Performance comparison for sequence descriptors and statistical algorithms

Feature type	Position	Composition	ACC	Thermo	3D	All
NVAR	84 (15)	84 (29)	108 (37)	36 (14)	330 (53)	642 (148)
SVM	0.64 (0.57)	0.60 (0.47)	0.47 (0.45)	0.53 (0.57)	0.53 (0.57)	0.71 (0.80)
PLS	0.64 (0.59)	0.55 (0.38)	0.41 (0.40)	0.55 (0.54)	0.56 (0.56)	0.69 (0.73)
RF	0.65 (0.56)	0.47 (0.39)	0.48 (0.48)	0.56 (0.56)	0.55 (0.56)	0.63 (0.66)

Correlation coefficients for SVM, PLS and RF for models trained with all variables or a reduce set (within brackets) are shown. Selection was performed on the entire variable space. The NVAR row lists the number of variables broken down by descriptor blocks, before and after the selection had been done. Each model combination was generated based on the Huesken training set of 2182 siRNAs, and the correlation coefficients are related to the predictions of the Huesken test set of 249 siRNAs. The same training and test sets were used as benchmarks from other publicly available predictors and their performance is reported here for comparison: $R^{BIOPREDsi[14]} = 0.66$, $R^{DSIR[15]} = 0.67$ and $R^{Thermocomposition[16]} = 0.66$.

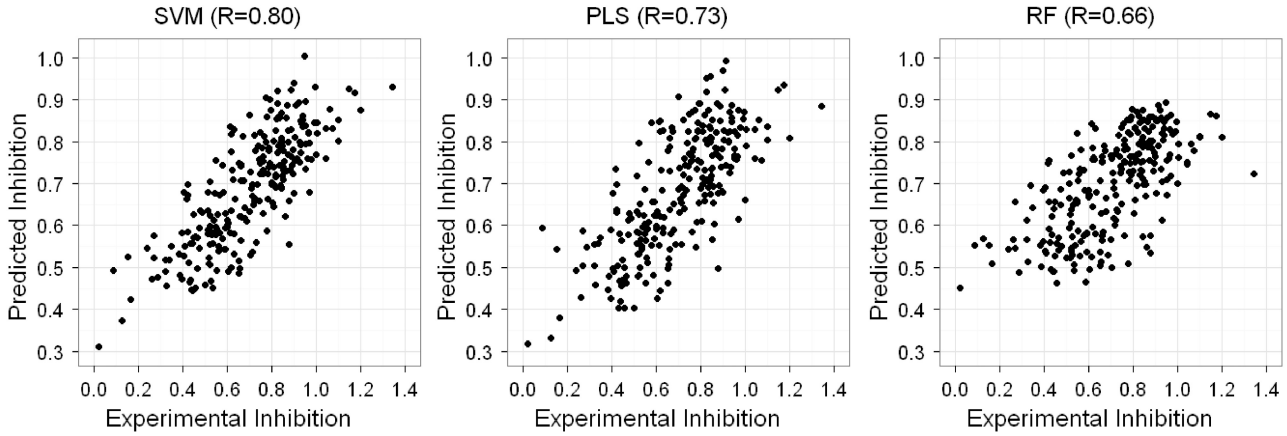

Figure 2. SVM, PLS and RF model predictions versus experimental siRNA activity for the test set of 249 sequences taken from the Huesken data set.

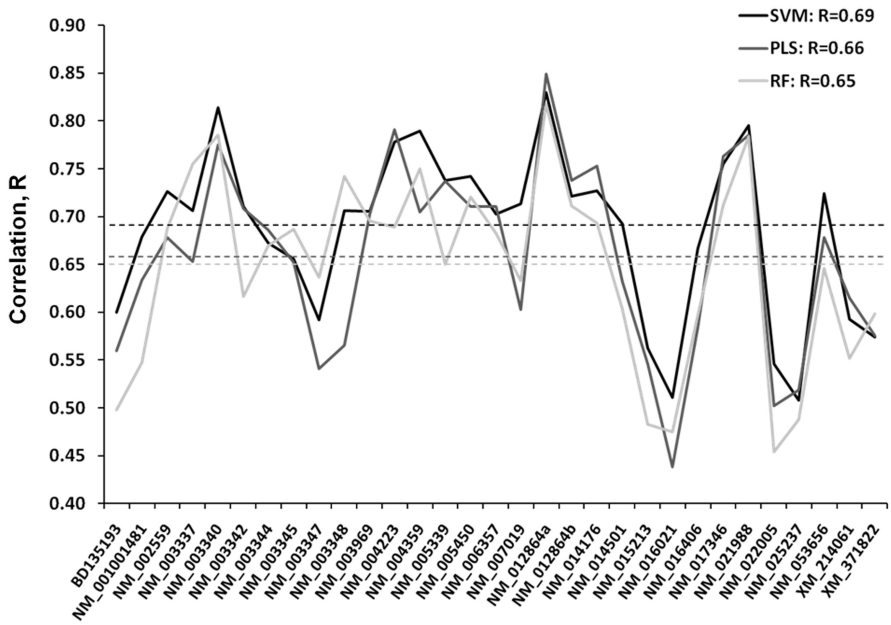

Figure 3. LOGO cross-validation results based on 31 gene sets from the Huesken data set (*x*-axis). Although prediction accuracy varies significantly with target mRNA, SVM showed overall the best correlation with experimental activity (*R*: >0.65 in 23 of 31 LOGO calculations).

An alternative cross-validation study was also run using the Huesken data in a strategy that can be described as Leave-One-Gene-Out (LOGO). This is a realistic scenario in a drug design setting, where an *in silico* model based on available data is applied to predict the activity of siRNA sequences derived from a new and unrelated gene target. Beyond being unbiased, the LOGO validation study provides an overall estimation of target dependency, which has not been fully addressed in the literature. The data set of 2431 siRNA sequences was encoded by the entire set of 642 descriptors, and in sequential rounds of analysis, the following protocol was used: (i) siRNAs for one gene were removed from the complete set to create a test set, (ii) multiple models were built (SVM, PLS and RF) from the remaining data (training set) and (iii) the reduced model was used to predict the activities of the siRNA sequences from the gene not included in the training set. Figure 3 shows the correlation coefficient profiles for SVM, PLS and RF across the 31 validation studies carried out by the LOGO procedure. Overall, SVM showed a better correlation between predicted and experimental values when compared with PLS and RF. Although the average correlation coefficients are in line with the validation study performed on the multi-gene test set, prediction accuracy varies significantly with target (Figure 3). This target dependence may be due to a number of factors, including mRNA turnover rate, accessibility of the target sequence to the RNAi machinery, secondary and tertiary structure of the mRNA, binding proteins and subcellular localization of the transcript, which all together may hamper or even prevent access of activated RISC to the corresponding target sequences,

leading to reduced silencing or completely blocking silencing (50,51). Transferability between genes is critical for an efficacy model, and the SVM scoring scheme gave a correlation coefficient greater than 0.65 for 23 of 31 genes, whereas PLS and RF achieved such a threshold for only 18 and 17 genes, respectively. Based on these results, the SVM regression algorithm coupled with a reduced set of 148 sequence descriptors was selected for use going forward and it has been implemented in our current siRNA design workflow for potency predictions (H. Xi *et al*., unpublished result).

To further validate the performance of the SVM model, it was tested against a series of internal and external data sets. Unfortunately, few large consistently assayed data sets are available for siRNA model building and validation. We therefore created an internal data set consisting of small RNAs designed against three genes (HIF1A, HK2 and HPSE), with the aim of removing questions of assay and target variation. For each target, 100 sequences were selected using a basic selection funnel designed to generate siRNA of potential therapeutic interest. The criteria were: (i) select sequences in target exons avoiding exon boundaries and the untranslated region; (ii) bias for selection of sequences with cross-species homology to allow a single compound to advance through rodent, non-human primate and human trials; (iii) avoid single nucleotide polymorphisms in the target sequence; (iv) prevent matches in the siRNA seed region to known miRNA seed sequences; (v) exclude immune stimulatory and G-tetrad forming motifs; (vi) minimize high complementary matches to off-target transcripts; (vii) sort the remaining sequences using the SVM model for siRNA efficiency described previously and (viii) introduce any chemical modifications to the siRNA. Although straightforward, this filtering and selection process can prove logistically challenging. At Pfizer, this was solved by creating an informatics platform for siRNA and antisense oligonucleotides design. This tool has recently been made available to the scientific community as a server and design client, with all of the code also being available as an Open Source project (H. Xi *et al*., unpublished result). Each of the 300 selected sequences was designed as a standard 21-mer siRNA containing target matching overhangs (2 nucleotides in length) on the sense and antisense strands, but also as right (R-Dicer) and left (L-Dicer) dicer substrates, using the conventions described by Rossi and colleagues (52). The design patterns are shown in Figure 4. This resulted in a data set of 891 compounds.

The assay results for the 300 21-mer siRNA are given in Figure 5 and Supplementary Table S4. All the sequences were tested for percent knockdown (KD) of the corresponding target using three different concentrations (0.08 nM, 0.4 nM and 2 nM). For a subset of sequences, additional data were collected at 10 nM. All measurements were done 24 h post-transfection using a branched DNA assay. Although they were all designed to be active, siRNAs were selected with a range of potencies for each of the three targets. Slightly more active compounds were generated for HPSE than the other two targets. The predictions for this set of compounds based on the SVM model described earlier are shown in Figure 6 as a receiver operator characteristic plot. The SVM model was used for classification, with a cut-off of 70% KD at 10 nM concentration being considered active. Overall the performance is significantly better than random for each of the targets, with approximately 50% of the siRNA selected being active. As seen in the LOGO study of the Huesken data set, the performance for each gene does vary, with the HIF-1α predictions being better than those for HPSE and HK2. However, this trend in the data is not correlated with the absolute potency of the compounds, as there are a number of HPSE compounds with greater overall potency. In contrast to the Huesken data set, which was randomly assembled, the compounds in this set are biased towards active motifs and avoid sequence traits that reduce activity.

The results for the Dicer substrate compounds are also reported in Figure 5 and Supplementary Table S4. Overall, the 21-mer siRNAs are slightly more potent than the Dicer design, with the R-Dicer generally more active than the L-Dicer at the 2-nM screening concentration. However, comparison of the absolute KD values between different length oligonucleotides is challenging, as transfection efficiency can differ with the length of the compound. This difficulty extends beyond the comparison of data to the predictions of siRNA activity for compounds with different design patterns and the creation of models from non-uniform data sets. Models built with inconsistent assay formats, design patterns or screening concentrations have an additional challenge in identifying

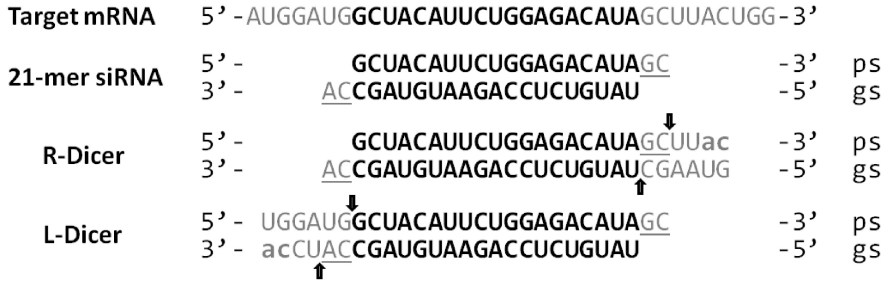

Figure 4. Sequence design for 21-mer siRNAs and R-/L-Dicer substrates. siRNAs were designed to have a stem length of 19-nts (nucleotides in bold) with 2-nts overhangs at each side matching the mRNA target (underlined nucleotides). The R-/L-Dicer substrates were designed based on the cleavage efficiency criteria defined by Rose *et al*., (53) where the Dicer entry is from the 2-nt 3′-overhang. The arrow points to the predicted cleavage position, and lowercase letters are used to indicate nucleotides with a 2′-deoxyribose sugar.

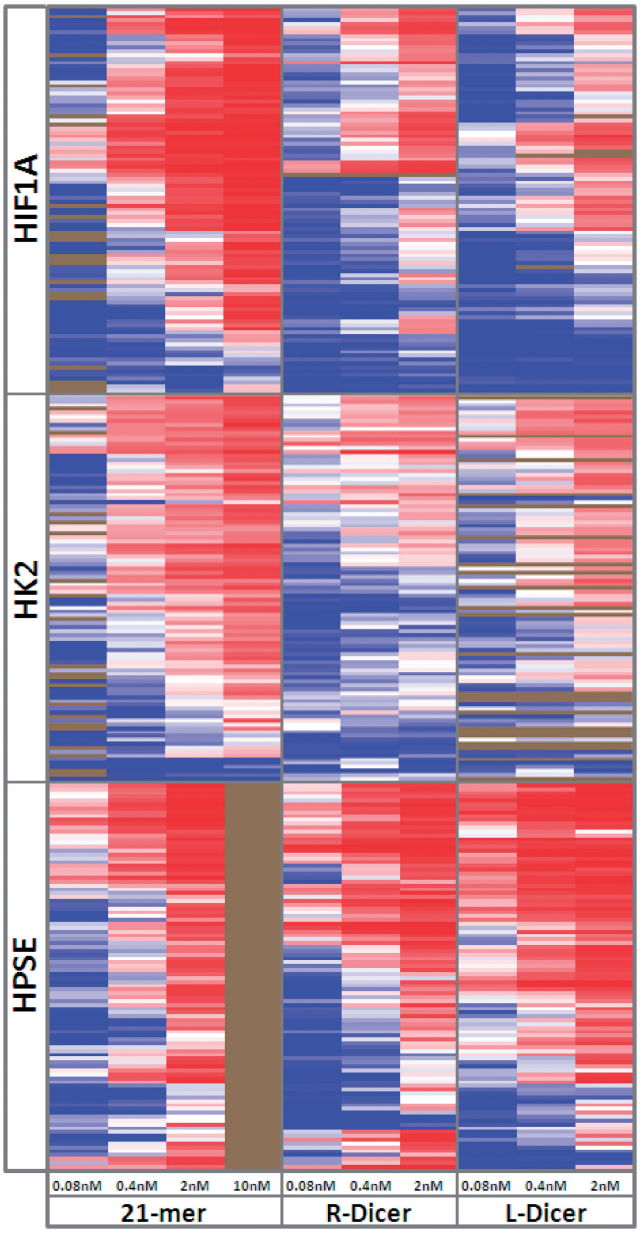

Figure 5. Eight hundred and ninety-one RNA sequences were designed against HIF1A, HK2 and HPSE as 21-mer, R-Dicer and L-Dicer designs. Sequences were tested at 0.08 nM, 0.4 nM, 2 nM and, in some cases, 10 nM. The 21-mer design showed the best hit rate (active: %KD ≥70, inactive: %KD <70) at all doses compared with R- and L-Dicer substrates in both HIF1A and HK2. Whereas for HIF1A, the R-design performed better than the L-design, the opposite was true for HK2. HPSE hit rates were higher for R-Dicer compared with 21-mer and L-Dicer, with 32 R-Dicer sequences showing >70% inhibition at 0.08 nM, versus 17 (L-Dicer) and 5 (21-mer). Sequence design and target mRNA both influence activity, with no clear trend observable.

features key to activity. This is manifested in a decreased prediction rate, as is seen in the dicer substrate data, where the features necessary for efficient dicer cleavage likely differ from those necessary to incorporate a 21-mer siRNA into the RISC complex.

In a final prospective study, the variability of model performance was examined across a wider set of targets.

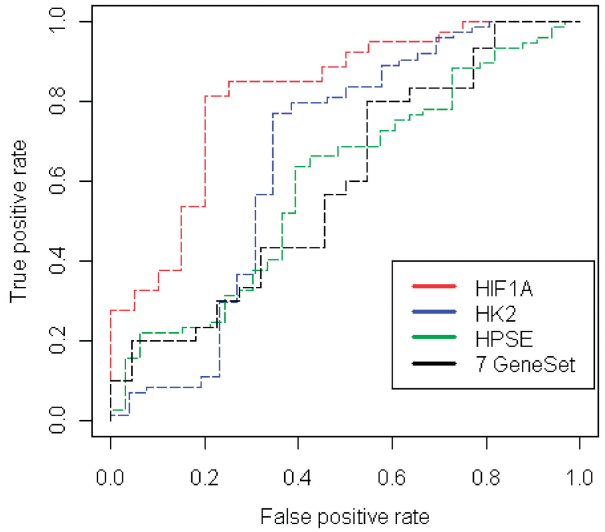

Figure 6. SVM model performance in the ROC space for the in-house prospective studies. The results for HIF1A showed the best predictive power among the studies genes (~70% accuracy at 40% false-positive rate). A slight decrease in the prediction power was observed for HK2 and HPSE, whereas the results for the 7 GeneSet lie on the random guess line (~50% accuracy at 40% false-positive rate).

Seven additional genes were selected (Survivin, C-Myc, EXH2, FRAP1, CTNB1, bRAF and PIK3Ca), which, like the initial three, are all relevant for their effect on HCC (7 GeneSet). The Ensembl ascension numbers for each gene are listed in the Supplementary Material along with the sequences and activities of the designed siRNA (Supplementary Data S1 and S4, respectively). Sequences were chosen using the workflow and filters described previously, with the compounds selected for synthesis being those predicted to be most active after filtering. In a receiver operator characteristic plot of the results (Figure 6), activity was defined as 70% KD at 10 nM. Using this cut-off, nearly half of the selected compounds are active when the data for the seven targets are combined, which is in line with the larger three-gene study. However, once separated by gene, the model performance is extremely target dependent, as can be seen in Figure 7. For example, nine of nine siRNAs designed for Survivin show an experimental KD of >70%. In contrast, for c-Myc, only one of the nine compounds shows a KD of >70%. This difference is not easily explained by examining the genes themselves, as they are approximately the same length with no unusual features, and although we attribute the difference in performance to the model, certain genes have proven to be much more difficult to knockdown with any siRNA. Larsson *et al.* (51) have recently investigated the importance of the mRNA turnover rate as a potential factor influencing mRNA susceptibility to perturbation by small RNA molecules, showing strong evidence for the inverse relationship between mRNA decay rates and propensity to siRNA-mediated gene silencing on a genome-wide scale. Their main conclusion was that real-world high-turnover transcripts were found to be more resistant to siRNA silencing, possibly explaining the difficulties we have found on knocking down the c-Myc gene, which is

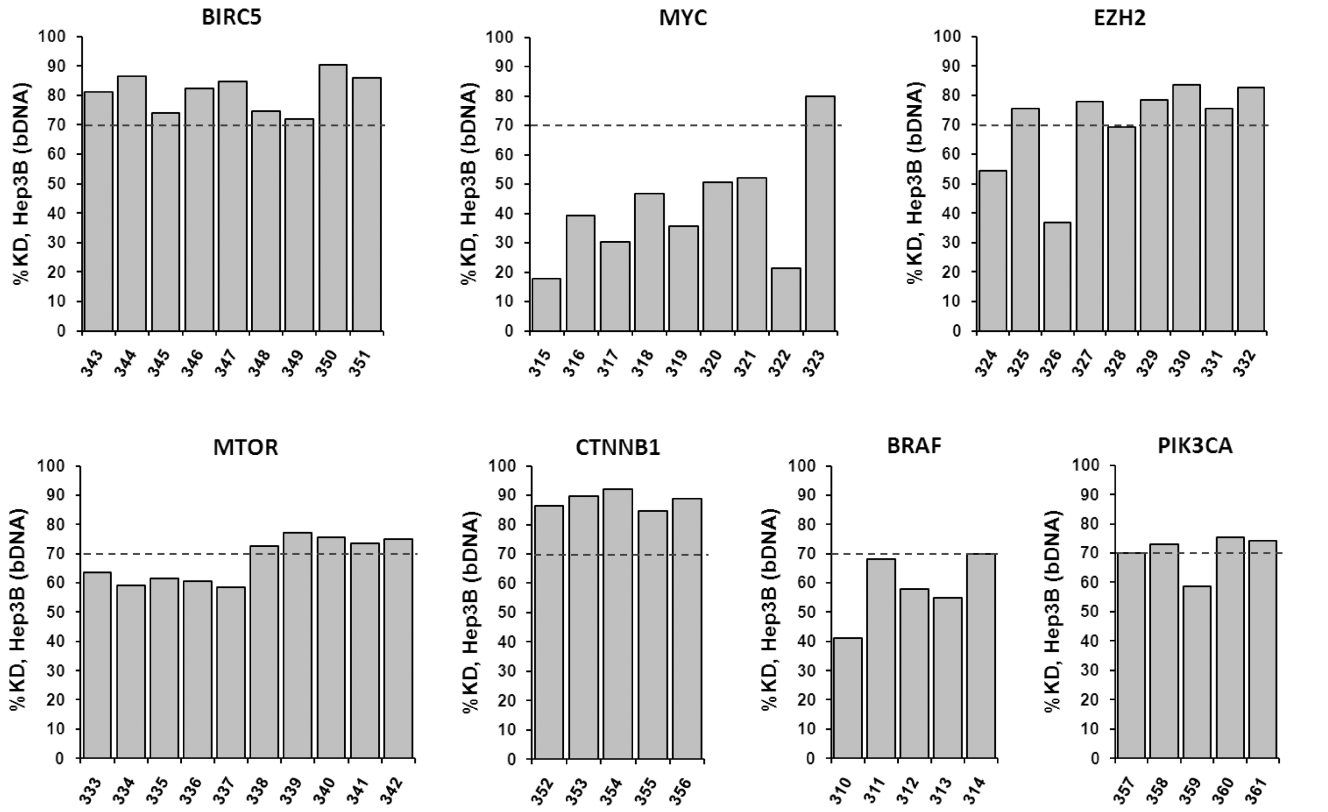

Figure 7. Rationally designed siRNAs targeting the mRNA of seven genes (7 GeneSet). All siRNAs were evaluated for silencing efficiency by measuring mRNA levels 24 h after transfection at 10-nM concentration. Model performance by gene shows a strong target dependency with cases like BIRC5 and CTNNB1, where all the designed sequences passed the 70% knockdown cut-off, whereas only one of five and one of nine active siRNAs were found for BRAF and MYC, respectively.

known for having an unusually high rate of cytoplasmic transcript turnover (c-Myc mRNA half-life $\leq$30 min) (53).

A true understanding of the relative rate of prediction for the model between genes would require each possible siRNA sequence within a gene to be tested and the total proportion of actual inactive siRNAs determined. This would then allow a true enrichment rate over random selection for each gene to be computed. Unfortunately, this is not feasible, and the typical use case more closely resembles the current study, where only a handful of siRNA will be made for any one target. The results for these seven genes show that for the majority of targets, a reasonable tool siRNA can be generated with a limited number of compounds; however, certain targets are likely to be more difficult.

To allow the performance of the SVM model generated here to be compared easily with previously published models, it was also profiled against three widely used data sets. The data sets studied were from Reynolds (244 siRNA targeting seven genes) (13), Vickers (76 siRNA targeting two genes) (25) and Harboth (44 siRNA targeting one gene) (26). These same data sets were used by Saetrom and Snove (54) and Ichihara *et al.* (17) to compare a wide variety of published models (Table 2). For the Reynolds data set, the SVM model has an R^2 of 0.54, which is consistent with the best of the other available models, DSIR (15) 0.54, i-score (17) 0.54 and GPboost (54) 0.55. For the Vickers data set, the results are similar, with an R^2 of 0.52 for the SVM model, comparing well with the models with the highest R^2s, including Ui-Tei *et al.* (12) 0.58, and i-score 0.58. For the smaller Harboth data set, the SVM model performed better than the other models, with an R^2 of 0.54 as compared with 0.51 for DSIR and 0.43 for Biopredsi. Unlike many of the other available models, the SVM model produces fairly consistent results across each of the external validation data sets.

CONCLUSION

Predicting amenable sites for RNAi intervention along the mRNA sequence of a target gene has been the focus of recent experimental and computational biology efforts. The finding that not all regions of a gene can be used to effectively trigger specific mRNA degradation lead to the investigation of statistical algorithms based on diverse nucleic acid descriptors for the selection of efficient and selective siRNA molecules. Although several factors were previously identified, which seem to contribute to the efficacy of siRNAs (thermodynamics of terminal duplex stability, preference of specific nucleotides at given positions in the duplex, sequence motifs recognition and sense/antisense competing reactions), it has become clear that these features do not provide an exhaustive description of the key determinants of siRNA potency. In this article, we described the application of 3D descriptors to

Table 2. Results of publicly available prediction algorithms on three independent test sets

Algorithm	$R^{Reynolds}$ (244si/7 g)	$R^{Vickers}$ (76si/2 g)	$R^{Harborth}$ (44si/1 g)
GPboost[a]	0.55	0.35	0.43
Ui-Tei[a]	0.47	0.58	0.31
Amarzguioui[a]	0.45	0.47	0.34
Hsieh[a]	0.03	0.15	0.17
Takasaki[a]	0.03	0.25	0.01
Reynolds 1[a]	0.35	0.47	0.23
Reynolds 2[a]	0.37	0.44	0.23
Schwarz[a]	0.29	0.35	0.01
Khvorova[a]	0.15	0.19	0.11
Stockholm 1[a]	0.05	0.18	0.28
Stockholm 2[a]	0.00	0.15	0.41
Tree[a]	0.11	0.43	0.06
Luo[a]	0.33	0.27	0.40
Ichihara (i-score)[b]	0.54	0.58	0.43
Huesken (Biopredsi)[b]	0.53	0.57	0.43
DSIR[b]	0.54	0.49	0.51
Katoh[b]	0.40	0.43	0.44
SVM[c]	0.54	0.52	0.54

[a]Saetrom and Snove (54).
[b]Ichihara *et al.* (17).
[c]SVM-based algorithm developed in this article. The model shows a stable high performance across the three external validation set.

capture RNA strain and flexibility, which have been shown to play an important role during the reversible adsorption of the antisense strand into RISC. These parameters describe near-equilibrium geometric deformations for RNA duplexes and were derived from MD simulations using state-of-the-art simulation conditions and last-generation force fields. Although the statistical performance of a model based only on 3D descriptors is on the same level of accuracy as models built on any other individual descriptor block, when combined together with other types of descriptors, an overall increase in model performance was observed. This can be explained by the variance in each block contributing independently with siRNA efficacy data. Among the statistical algorithms studied, the kernel regularized approach SVM consistently outperformed both linear (PLS) and non-linear (RF) regression techniques, and was chosen for the final model, which has been made available through the PFRED Open Source project.

The final model including 3D information shows equivalent or better performance in comparison with previously published algorithms, with increased consistency across multiple data sets; however, the fraction of sequences that do not translate from potent *in silico* siRNAs to *in vitro* functionally active siRNAs is still considerable for all of the available siRNA models. This suggests that further improvements may require a more fundamental understanding of the RNAi pathway at the molecular level. Obviously other phenomena, which have not been included in this study, such as hybridization kinetics, sense versus antisense competing reactions, mRNA turnover rate, local target accessibility and off-target effects may also need to be considered in a single workflow for a more accurate design. Room for improvement is also possible related to the extension of the prediction algorithm to incorporate the use of modified oligonucleotides. Our methodology is flexible enough as to allow an easy integration of non-coding nucleotides with a small cost of parameterization (mostly derivation of experimental stability descriptors, and a few MD simulations for calibration), which is where our efforts are currently focused.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–4.

ACKNOWLEDGEMENTS

The authors thank Shobha Potluri for preliminary discussions on the development of the algorithm, Simon Hualin and Christine Lawrence for the algorithm porting in PFRED, Jeremy Little for the synthesis of the siRNA and Xu Xu and Diana Gikunju for the biological data of the siRNAs included in this manuscript. M.O. thanks the support of the Spanish Ministry of Innovation and Science (BIO2009-10964, BIO2012-32868 and Consolider E-Science), Instituto Nacional de Bioinformática, Fundación Marcelino Botín and the ERC for an Advanced Grant.

FUNDING

Funding for open access charge: Pfizer.

Conflict of interest statement. None declared.

REFERENCES

1. Robinson,R. (2004) RNAi therapeutics: how likely, how soon? *PLoS Biol.*, **2**, 18–20.
2. Sudarsana,L.R., Sarojamma,V. and Ramakrishna,V. (2007) Future of RNAi in medicine: a review. *World J. Med. Sci.*, **2**, 1–14.
3. Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.
4. Hannon,G.J. and Rossi,J.J. (2004) Unlocking the potential of the human genome with RNA interference. *Nature*, **431**, 371–378.
5. Park,J.-E., Heo,I., Tian,Y., Simanshu,D.K., Chang,H., Jee,D., Patel,D.J. and Kim,V.N. (2011) Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature*, **475**, 201–205.
6. Meister,G., Landthaler,M., Patkaniowska,A., Dorsett,Y., Teng,G. and Tuschl,T. (2004) Human argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell*, **15**, 185–197.
7. Peters,L. and Meister,G. (2007) Argonaute proteins: mediators of RNA silencing. *Mol. Cell*, **26**, 611–623.
8. Liu,J., Carmell,M.A., Rivas,F.V., Marsden,C.G., Thomson,J.M., Song,J.-J., Hammond,S.M., Joshua-Tor,L. and Hannon,G.J. (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science (New York, NY)*, **305**, 1437–1441.
9. Tuschl,T., Zamore,P., Lehmann,R., Bartel,D. and Sharp,P. (1999) Targeted mRNA degradation by double-stranded RNA *in vitro*. *Genes Dev.*, **13**, 3191–3197.
10. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
11. Amarzguioui,M. and Prydz,H. (2004) An algorithm for selection of functional siRNA sequences. *Biochem. Biophys. Res. Commun.*, **316**, 1050–1058.
12. Ui-Tei,K., Naito,Y., Takahashi,F., Haraguchi,T., Ohki-Hamazaki,H., Juni,A., Ueda,R. and Saigo,K. (2004)

Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.*, **32**, 936–948.
13. Reynolds,A., Leake,D., Boese,Q., Scaringe,S., Marshall,W.S. and Khvorova,A. (2004) Rational siRNA design for RNA interference. *Nat. Biotech.*, **22**, 326–330.
14. Huesken,D., Lange,J., Mickanin,C., Weiler,J., Asselbergs,F., Warner,J., Meloon,B., Engel,S., Rosenberg,A., Cohen,D. *et al.* (2005) Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotech.*, **23**, 995–1001.
15. Vert,J.-P., Foveau,N., Lajaunie,C. and Vandenbrouck,Y. (2006) An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics*, **7**, 520.
16. Shabalina,S., Spiridonov,A. and Ogurtsov,A. (2006) Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics*, **7**, 65.
17. Ichihara,M., Murakumo,Y., Masuda,A., Matsuura,T., Asai,N., Jijiwa,M., Ishida,M., Shinmi,J., Yatsuya,H., Qiao,S. *et al.* (2007) Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. *Nucleic Acids Res.*, **35**, e123.
18. Hofacker,I., Fontana,W., Stadler,P., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
19. Lu,Z.J. and Mathews,D.H. (2008) Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res.*, **36**, 640–647.
20. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
21. Wang,Y., Sheng,G., Juranek,S., Tuschl,T. and Patel,D.J. (2008) Structure of the guide-strand-containing argonaute silencing complex. *Nature*, **456**, 209–213.
22. Wold,S., Jonsson,J., Sjiistriim,M., Sandberg,M. and Rannar,S. (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta*, **217**, 239–253.
23. Faustino,I., Perez,A. and Orozco,M. (2010) Toward a consensus view of duplex RNA flexibility. *Biophys. J.*, **99**, 1876–1885.
24. Huesken,D., Asselbergs,F., Kinzel,B., Natt,F., Weiler,J., Martin,P., Haner,R. and Hall,J. (2003) mRNA fusion constructs serve in a general cell-based assay to profile oligonucleotide activity. *Nucleic Acids Res.*, **31**, e102.
25. Vickers,T.A., Koo,S., Bennett,C.F., Crooke,S.T., Dean,N.M. and Baker,B.F. (2003) Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *J. Biol. Chem.*, **278**, 7108–7118.
26. Harborth,J., Elbashir,S.M., Vandenburgh,K., Manninga,H., Scaringe,S.A., Weber,K. and Tuschl,T. (2003) 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**, 83–105.
27. Wincott,F., DiRenzo,A., Shaffer,C., Grimm,S., Tracz,D., Workman,C., Sweedler,D., Gonzalez,C., Scaringe,S. and Usman,N. (1995) Synthesis, deprotection, analysis and purification of RNA and ribosomes. *Nucleic Acids Res.*, **23**, 2677–2684.
28. Schwarz,D.S., Hutvagner,G., Du,T., Xu,Z., Aronin,N. and Zamore,P.D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.
29. Xia,T., SantaLucia,J., Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base Pairs. *Biochemistry*, **37**, 14719–14735.
30. Cornell,W.D., Cieplak,P., Bayly,C.I., Gould,I.R., Merz,K.M., Ferguson,D.M., Spellmeyer,D.C., Fox,T., Caldwell,J.W. and Kollman,P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.
31. Perez,A., Marchan,I., Svozil,D., Sponer,J., Cheatham,T.E. III, Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
32. Pérez,A., Blas,J.R., Rueda,M., López-Bes,J.M., de la Cruz,X. and Orozco,M. (2005) Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.*, **1**, 790–800.
33. Perez,A., Luque,F.J. and Orozco,M. (2007) Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.*, **129**, 14739–14745.
34. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
35. Sherer,E.C., Harris,S.A., Soliva,R., Orozco,M. and Laughton,C.A. (1999) Molecular dynamics studies of DNA A-tract structure and flexibility. *J. Am. Chem. Soc.*, **121**, 5981–5991.
36. Strahs,D. and Schlick,T. (2000) A-tract bending: insights into experimental structures by computational models. *J. Mol. Biol.*, **301**, 643–663.
37. Kretschmer-Kazemi Far,R. and Sczakiel,G. (2003) The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucleic Acids Res.*, **31**, 4417–4424.
38. Patzel,V., Rutz,S., Dietrich,I., Koberle,C., Scheffold,A. and Kaufmann,S.H.E. (2005) Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency. *Nat. Biotech.*, **23**, 1440–1444.
39. Kurreck,J. (2006) siRNA efficiency: structure or sequence—that is the question. *J. Biomed. Biotechnol.*, **2006**, 1–7; doi: 10.1155/JBB/2006/83757.
40. Heale,B.S.E., Soifer,H.S., Bowers,C. and Rossi,J.J. (2005) siRNA target site secondary structure predictions using local stable substructures. *Nucleic Acids Res.*, **33**, e30.
41. Gredell,J.A., Berger,A.K. and Walton,S.P. (2008) Impact of target mRNA structure on siRNA silencing efficiency: a large-scale study. *Biotechnol. Bioeng.*, **100**, 744–755.
42. Tafer,H., Ameres,S.L., Obernosterer,G., Gebeshuber,C.A., Schroeder,R., Martinez,J. and Hofacker,I.L. (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotech.*, **26**, 578–583.
43. Eriksson,L., Antti,H., Gottfries,J., Holmes,E., Johansson,E., Lindgren,F., Long,I., Lundstedt,T., Trygg,J. and Wold,S. (2004) Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.*, **380**, 419–429.
44. Mevik,B.-H. and Wehrens,R. (2007) The pls package: principal component and partial least squares regression in R. *J. Stat. Softw.*, **18.**
45. Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, **2**, 121–167.
46. Vapnik,V. (2000) Information science and statistics. In: Jordan,M., Nowak,R. and Schölkopf,B. (eds), *The Nature of Statistical Learning Theory*, Vol. 19. Springer, Germany, p. 48.
47. Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.
48. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
49. Baroni,M., Costantino,G., Cruciani,G., Riganelli,D., Valigi,R. and Clementi,S. (1993) Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct. Act. Relat.*, **12**, 9–20.
50. Krueger,U., Bergauer,T., Kaufmann,B., Wolter,I., Pilk,S., Heider-Fabian,M., Kirch,S., Artz-Oppitz,C., Isselhorst,M. and Konrad,J. (2007) Insights into effective RNAi gained from large-scale siRNA validation screening. *Oligonucleotides*, **17**, 237–250.
51. Larsson,E., Sander,C. and Marks,D. (2010) mRNA turnover rate limits siRNA and microRNA efficacy. *Mol. Syst. Biol.*, **6**, 433.
52. Rose,S.D., Kim,D.-H., Amarzguioui,M., Heidel,J.D., Collingwood,M.A., Davis,M.E., Rossi,J.J. and Behlke,M.A. (2005) Functional polarity is introduced by Dicer processing of short substrate RNAs. *Nucleic Acids Res.*, **33**, 4140–4156.
53. Jones,T.R. and Cole,M.D. (1987) Rapid cytoplasmic turnover of c-myc mRNA: requirement of the 3′ untranslated sequences. *Mol. Cell. Biol.*, **7**, 4513–4521.
54. Saetrom,P. and Snove,J.O. (2004) A comparison of siRNA efficacy predictors. *Biochem. Biophys. Res. Commun.*, **321**, 247–253.