# A comparison of siRNA activity predictors using advanced regression techniques

Simone Sciabola, Qing Cao, Theresa Johnson, Lingling Shen, Robert Stanton,
Xiaoyu Jiang, Simon Xi, Jason Hughes, Daniel Caffrey, Shobha Potluri, Steven Haney, Johnathan Cyr, Jeremy Little (Pfizer RTC, Cambridge, MA)

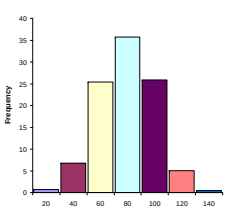
## Introduction

- RNA interference (RNAi) has been called "one of the most exciting discoveries in biology in the last couple decades" and it has rapidly become one of the most powerful tools in the field of functional genomics by enabling genome-scale loss of function screens in cultured cells. Recently there has been a rapid progress towards its use as therapeutic modality against human diseases.
- Short interfering (siRNA) are duplex RNA strands which are incorporated into the RNA induced silencing complex (RISC) and degrade complementary messenger RNA (mRNA) sequences. However, previous studies indicate that not all the siRNAs produce the same knock down effects. Therefore a key component of RNAi applications is the selection of effective siRNA sequences which are highly efficient in degrading target mRNA.
- Although considerable progress has been made recently in understanding how gene silencing is mediated by the RNAi pathway, the design of effective sequences is still a challenging task. To tackle the efficacy challenge and further improve accuracy for prediction of siRNA potency, we performed a comparison of advanced regression techniques (SVM, GPR, PLS and LASSO) and new in-house generated descriptors (sequence position, sequence compositions, ACC transform, thermodynamic and secondary structure descriptors) in the generation of siRNA efficiency models.

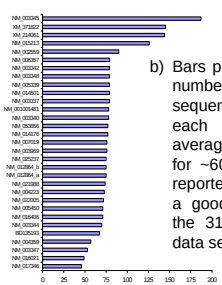
## Material and Methods

### Data collection

- Accurate experimental data is a prerequisite for any reliable QSAR model. Although several siRNA data sets are now publicly available[1-6], care must be taken in the way they are merged together and used as input data for statistical model building.
- Some of the potential issues associated with these data sets are: (i) a variety of assays for measurement of siRNA efficacy; (ii) different siRNA concentrations; (iii) sub-optimal intervals between transfection and down-regulation measurement; (iv) biased introduced in the selection of both target genes and siRNA sequences.
- In this respect the publication in Huesken et al [7] of an unbiased set of 2431 randomly selected siRNAs targeting 31 mRNA constructs assayed through the same high-throughput fluorescent reporter gene system represents a milestone, and this data was used in our study (Figure a-b).



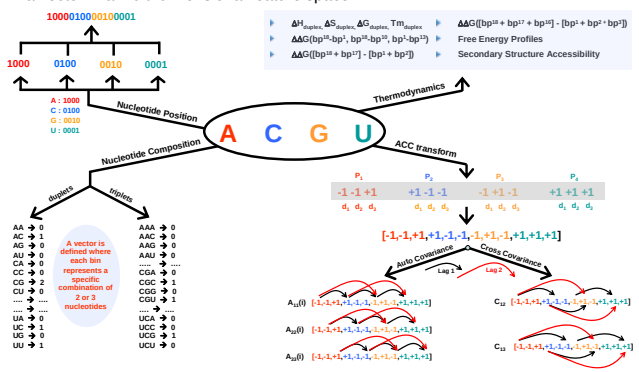

a) Density distribution of inhibitory activities of 2431 randomly selected siRNAs targeting 31 reporter plasmids.




b) Bars plot showing the number of siRNA sequences tested for each gene; on average the inhibition for ~60 siRNAs were reported representing a good coverage for the 31 genes in the data set.

### siRNA sequence description

- In order to develop any statistical model for siRNA efficacy prediction we must first choose a set of numerical features to represent a given oligonucleotide sequence as a vector in a multi-dimensional feature space.




- The final descriptors matrix was a combination of the four different class of numerical features.

### Statistical Algorithms

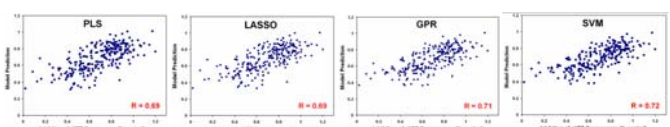
We then subjected the feature vector representing training sequences to a variety of supervised regression algorithms. Indeed, regression analysis is one of the most widely used statistical techniques for prediction and it takes various forms, including standard linear regression, kernel regularized learning and Bayesian approaches. The following were implemented and tested in our study:

- **LASSO** (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) regression searches for the optimal coefficient estimates in a restricted space by setting less important variable coefficients to zero.
- **PLS** (**P**artial **L**east **S**quares) regression finds the multidimensional directions in the predictor space that explains the most variance direction in the dependent variable space.
- **SVM** (**S**upport **V**ector **M**achine) learns the non-linear pattern in the training data by a linear learning machine in a kernel-induced space and produces the predictions for the test data.
- **GPR** (**G**aussian **P**rocess **R**egression) is a Bayesian technique which draws the posterior information of predictions by a Gaussian process specified by the data and prior information.

## Results

### Model building

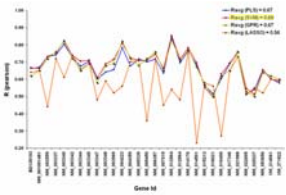
- All the models where calibrated using the Huesken training set 2182 siRNA sequences and validated on the Huesken test set (249 randomly picked siRNAs). The predictions for the test set are graphically represented below, where each box shows the correlation between predicted and observed inhibitory activities obtained with different statistical algorithms:




- All methods performed in line with previously reported results for the same Novartis test set ($R^{DSIR[8]} = 0.67$, $R^{BIOPREDsi[7]} = 0.66$, $R^{Thermocomposition[5]} = 0.66$), with SVM and GPR being slightly of better-quality.

### Leave-One-Gene-Out (LOGO) internal validation

- Leave One Out (LOO) cross validation works by building reduced models, where one object at a time is removed, and using them to predict the Y of the object held out.
- To test whether the model could be reliably applied to predict new data, a more efficient validation procedure will consist on ***L***eaving ***O***ne ***G***ene at a time ***O***ut (***LOGO***). In this way, all the siRNAs targeting one gene are removed from the training set, and the model trained with the remaining siRNAs is used to predict the *"N"* siRNAs left out.



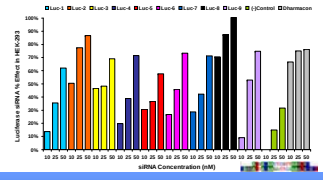

### External Validation

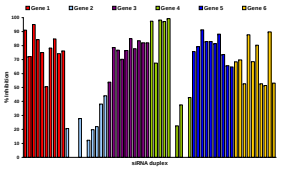
- To further assess the predictive power of our model (we chose **SVM** for this comparison) we tested it on five independent data set. Overall the SVM model performed equally well, in same cases better, when compared with other well known algorithms, although a clear drop in the performance between the Huesken test set and the independent test sets can be observed among all the reported methods.

Pearson coefficients recomputed from Ichihara (Nucleic Acids Research 2007, gkm498)

| Source | siRNAs | $R_{BIOPREDsi}$ | $R_{i-score}$ | $R_{Katoh}$ | $R_{SVM}$ | $R_{DSIR}$ |
|---|---|---|---|---|---|---|
| Reynolds | 244 (7) | 0.53 | 0.54 | 0.40 | **0.55** | 0.54 |
| Vickers | 76 (2) | 0.57 | 0.58 | 0.43 | **0.48** | 0.49 |
| Harborth | 44 (1) | 0.43 | 0.43 | 0.44 | **0.55** | 0.51 |
| Shabalina | 653 (52) | / | / | / | **0.47** | 0.48 |
| Katoh | 700 (1) | / | / | / | **0.57** | / |

- The SVM model was applied to rationally design siRNAs against the Firefly Luciferase gene (GeneBank no. U47298, pGL3 isoform). Top 10 predicted siRNAs (19N+dTdT) were test-ed in the Luciferase assay at 10, 25 and 50 nM siRNA concentration and Luc activity was measured 24h after transf-ection in HEK-293 cells (results shown below).
- siRNAs seqs (19N+UU) intended to trig-ger RNAi in 6 endogenously expressed mRNA transcripts were designed and the SVM model was applied to prioritize them. Top 10 predicted siRNAs were tested at 12.5 nM concentration and mRNA levels were measured 24h after transfection in MCF-7 cells by using Taqman assay (results shown below).




## Outlook and Future Work

- Overall, kernel regularized approaches (SVM, GPR) seem to perform better than linear regression techniques (PLS, LASSO), with GPR, SVM and PLS being consistently the most stable in terms of prediction power. The poor performance of LASSO in LOGO cross validation may be due to the correlated observations in the data, which violates the underlying assumption for LASSO, and LASSO's mechanism of leaving out correlated descriptors, which may cause relevant information loss.
- The roughly same predictions shown by some "state of the art" statistical algorithm such as PLS, SVM and GPR, also in line with previously reported results obtained by using different combination of numerical features and statistics, might tell us that in order to improve in predicting siRNA efficacy the work to do lies more in coming up with new Design Of Experiments and a better understanding of the inhibition pathway at the molecular level, than trying other methods.

## Literature

**1)** Vickers et al. Efficient Reduction of Target RNAs by Small Interfering RNA and RNase H-dependent Antisense Agents. A comparative analysis. *J. Biol. Chem.* 2003, 278, 7108-7118. **2)** Harborth 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* 2003, 13, 83-105. **3)** Khvorova et al. Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell* 2003, 115, 209-216. **4)** Reynolds et al. Rational siRNA design for RNA interference. *Nat Biotech* 2004, 22, 326-330. **5)** Shabalina et al. Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics* 2006, 7, 65. **6)** Katoh et al. Specific residues at every third position of siRNA shape its efficient RNAi activity. *NAR* 2007, 35, e27. **7)** Huesken et al. Design of a genome-wide siRNA library using an artificial neural network. *Nat Biotech* 2005, 23, 995-1001. **8)** Vert et al. An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* 2006, 7:520