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, Shohba 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.

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.

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 introduction 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 mRNAs 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).

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).

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 Leaving One Gene at a time Out (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 a set of five independent test sets, and the SVM model was applied to prioritize them. Top 10 predicted siRNAs were 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.

The SVM model was applied to rationally design siRNAs against the Firefly Luciferase gene (GeneBank no. U47296, pGL3 isofrom). Top 10 predicted siRNAs (19N+U+T+T) were test-ed in the Luciferase assay at 10, 25 and 50 nM siRNA concentration and Luc activity was measured 24h after transfection in HEK-293 cells (results shown below).

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 information of predictions by a Gaussian process specified by the data and prior information.

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.

References