

# Applying machine learning for rational siRNA design

Spiridonov A<sup>1</sup>, Ogurtsova A, Shabalina S<sup>2</sup>

**Keywords:** siRNA design, Support Vector Machines (SVM), thermodynamic profile, consensus

## 1 Introduction

RNA interference (RNAi) is a promising method to suppress gene expression in eukaryotic cells. Both micro RNAs (miRNAs) and short interfering RNAs (siRNAs) have been identified as sequence-specific posttranscriptional regulators of gene expression. It was shown that miRNAs and functional siRNAs exhibit strand bias, have common thermodynamic features, and most likely interact with the same set of cellular proteins [3].

This study focuses on building a SVM-based model of siRNA interference, driven by the following two motivations. Firstly, we can expedite experiments by ruling out bad siRNA candidates. Secondly, knowing which features have the greatest effect on efficiency, we can speculate about the mechanism of RNAi. Our goal was to find an optimal set of parameters for quick siRNA prediction which would allow creation of a database of effective siRNA targets for total human transcriptome.

## 2 Data and software

We calculated thermodynamic and composition features for a set of 653 siRNAs with known functional efficiencies from a number of diverse experiments using our own scripts. We used support vector machine (SVM)-based feature selection to extract key parameters from the resulting features. The selection was done using an iterated greedy approach: from the available features, add the ones that substantially improve performance; then, remove any features that do not help significantly; repeat until no changes occur.

Performance was evaluated using 7-fold cross-validation (CV) to produce a “blind” prediction for every point in the dataset. Since our data set contained a substantial number of overlapping sequences, we used a custom cross-validation scheme similar to minus-mRNA, which ensured that overlapping sequences ended up in the same CV part. We used the support vector machine implementation in LIBSVM 2.71 [1]. From the available algorithms, we chose nu-SVR (a variant of support vector regression). The training parameters C and gamma were optimized using a modified version of the LIBSVM grid-search script.

---

<sup>1</sup>Department of Applied Mathematics, 2-333 Massachusetts Institute of Technology, Cambridge, MA 02139 E-mail: lesha@mit.edu

<sup>2</sup>National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, Bethesda, MD, 20894 E-mail: shabalin@ncbi.nlm.nih.gov

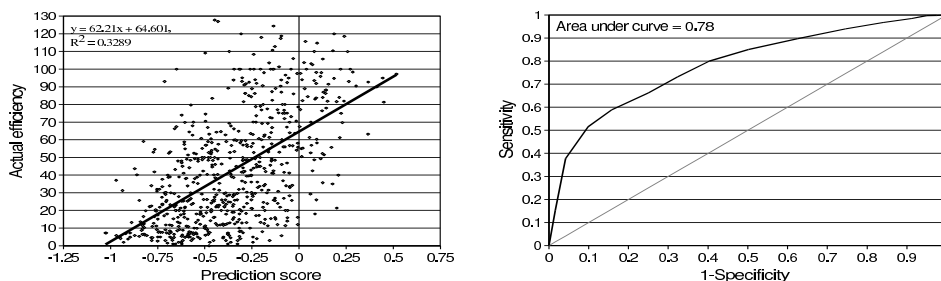


Figure 1: (left) Linear regression of cross-validation SVM score versus actual siRNA activity and (right) the ROC curve for SVM score-based threshold classification.

### 3 Results

For the experimentally studied siRNA-mRNA pairs, we calculated a number of thermodynamic and content parameters – some novel, others adopted from the literature. We found that some content indexes show the same trend in both miRNAs and efficient siRNAs. We also found significant correlations between some of the parameters and the silencing efficiency of siRNAs in our dataset of 653 sequences. It is well known that mRNA secondary structure influences siRNA efficiency [2]. However, we did not take such features into account, because calculating them is quite expensive on genome-scale datasets (quadratic time, or slower).

We used support vector machine (SVM)-based feature selection to extract 21 key parameters from our list of  $\approx 250$  features. An SVM model of siRNA activity using the resulting parameters had  $R^2 = 0.329$ , see Figure 1. Thus, our parameters using the siRNA sequence alone allow reasonable predictions. We also used an SVM score threshold (derived with a cross-validation procedure) to make classification decisions. The resulting ROC curve for classifying active (residual activity is under 35%) versus inactive siRNAs is in Figure 1. These key 21 prediction parameters were calculated for every human mRNA sequence in RefSeq database.

Our new set of thermodynamic and composition parameters and the machine learning methodology can be used to improve siRNA design in future genomic studies.

### References

- [1] Chih-Chung, Chang and Chih-Jen, Lin. 2001. LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [2] Heale, B.S., Soifer, H.S., Bowers, C., Rossi, J.J. 2005. siRNA target site secondary structure predictions using local stable substructures. *Nucleic Acids Res.* 33:e30.
- [3] Khvorova, A., Reynolds, A., Jayasena, S.D. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115:209-216.