Poster M-20
Optimized Selection of Short Interfering RNA Target Loci

Authors:
Istvan Ladunga (*Center for Biotechnology & Dept. Statistics, University of Nebraska-Lincoln*)

Short Abstract: To increase the silencing efficacy of short interfering RNA (siRNA), we developed optimized target selection methods. The constrained optimization, support vector machine and partial least squares methods were trained on 3000 siRNA experiments to predict silencing efficacy with an accuracy of 88 percent in cross-validation and blind-test experiments.

Long Abstract:
To increase the silencing efficacy of short interfering RNA (siRNA) molecules, we present novel, optimized target site selection methods. Our optimized target site selection tools are designed to achieve more complete silencing of the targeted gene, a critical issue in both the experimental assessment of protein function and in siRNA-based candidate therapeutics. This objective is best accomplished by targeting the minimal number of mRNA loci and applying low siRNA dose so as to both minimize silencing of untargeted mRNAs [1] and to avoid saturating the RNA-Induced Silencing Complex (RISC) [2]. Heuristic methods [3], and an undisclosed method of artificial neural networks [4] have been published to discriminate between low- and high efficacy siRNAs.
Support vector machines [5] predicted efficacies based on 315 artificial oligonucleotides. We present further improvements to the prediction of actual knockdown efficacy values using nine times more single siRNA experiments, a larger pool of initial features, and different artificial intelligence methods. From the largest series of experiments, performed at Novartis [4], we chose
2252 siRNAs or short hairpin RNAs targeted against 28 genes; from the Dharmacon set [3] we used 238 siRNA molecules directed to 8 genes; and from the AObase [6], we took 542 artificial
oligonucleotides targeted to 45 mammalian genes.
These datasets allowed us to develop and validate robust, iterative optimal feature selection and weighting methods that predict the silencing activity in the function of sequence, thermodynamic, accessibility and other features of both the siRNA duplex and its target. Sites with less than 5 mismatches to any other coding sequence were disregarded. We computed positional, regional and global free energy, enthalpy, entropy features on the basis of Xia et al.'s observations [7] and stacking energies as recommended by [8]. Probability distributions of the target accessibility were sampled by both the sfold [9, 10] and the RNAup [11] methods. We also calculated the distance between the translation initiating codon and the start of the target site; the contents of tri- and tetramers with identical nucleotides; positional, regional and global mono- and dinucleotide frequencies; and self-hairpin energy. All feature values were normalized so as to facilitate iterative elimination of features with low contribution to the predictions. Our prediction methods were trained, cross-validated and blind-tested on all the three sets of experimentally observed single-siRNA activity data. The heterogeneity of the low-efficacy siRNAs necessitates a two-tiered approach. In the first step,

a two-category misclassification minimization method classifies each siRNA into the low-, high-efficacy group or to the borderline category. Second, robust linear and nonlinear regression models were solved by constrained optimization [12], support vector machine [13] or partial least square algorithms [14]. The selected features in an optimally weighted combination reliably predict the knockdown efficacy including positional nucleotide and thermodynamic (free energy) preferences, the lack of self-hairpin formation, etc.

Some of these features represent positional sophistications of previously recommended overall G+C content and other global features. The previously reported thermodynamic profile signature [15], is now extended to positional dinucleotide enthalpy, enthalpy/entropy and sequence features in addition to change in free energy.

Using one-half of the data for training and the other for testing in ten cross-validation experiments indicated a classification accuracy of 88 percent between the low- and high-efficacy groups. For the high-efficacy (>60 percent) group, regression accuracies of 86 and 92 percent were achieved. We note that these predictions require some calibration of the methods for each untrained cell line/transfection agent combination. Fortunately, even if the numerical efficacy values differ between such conditions, the ranking of the target loci predicted for our conditions can still be used as a guideline. For organisms other than mammals, the methods will need to be retrained and retested when adequate number of single-siRNA experimental observations becomes available.

By the end of May 2006, we finish our web-server for predicting the activity of siRNA molecules, and make the executables of the prediction program freely available for academic researchers.

REFERENCES

1. Sarov M, Stewart AF: The best control for the specificity of RNAi. Trends Biotechn. 2005, 23:446-448.

2. Meister G, Tuschl T: Mechanisms of gene silencing by double-stranded RNA. Nature 2004, 431:343-349.

3. Reynolds A et al>: Rational siRNA design for RNA interference. Nat Biotech 2004, 22:326-330.

4. Huesken D et al: Design of a genome-wide siRNA library using an artificial neural network. Nat Biotechnol 2005, 23:995-1001.

5. Camps-Valls G, et al.: Profiled support vector machines for antisense oligonucleotide efficacy prediction. BMC Bioinformatics 2004, 5:135.

6. Bo X, Lou S, Sun D, Yang J, Wang S: AOBase: a database for antisense oligonucleotides selection and design. Nucleic Acids Res 2006, 34:D664-667.

7. Xia T et al.: Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 1998, 37:14719-14735.

8. Muller M: Statistical physics of RNA folding. Phys Rev E Stat Nonlin Soft Matter Phys 2003, 67:021914.

9. Ding Y, Chan CY, Lawrence CE: RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA 2005, 11:1157-1166.

10. Ding Y, Chan CY, Lawrence CE: Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res 2004, 32:W135-141.

11. Muckstein U et al.: Thermodynamics of RNA-RNA Binding. Bioinformatics 2006:l024.

12. Ladunga I: PHYSEAN: PHYsical SEquence ANalysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. Bioinformatics 1999, 15:1028-1038.
13. Scholkopf B, Smola A: Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond 2001.
14. de Jong S: SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 1993, 18:251-263.
15. Khvorova A et al.: Functional siRNAs and miRNAs exhibit strand bias. Cell 2003, 115:209-216.