A machine learning approach for identifying the PTEN non-coding ceRNA network

Swami Iyer$^{\alpha}$, Robert Moray$^{\beta}$, Prajna Kulkarni$^{\alpha}$, Rahul Kulkarni$^{\alpha}$ and Kourosh Zarringhalam$^{\beta}$
Department of Physics$^{\alpha}$ and Department of Mathematics$^{\beta}$
University of Massachusetts at Boston

Introduction

PTEN is one of the most commonly altered tumor-suppressor genes in human cancers. It has been shown that even a subtle decrease in PTEN levels can significantly increase tumor susceptibility, whereas elevation of PTEN levels can induce a tumor-suppressive metabolic state. Recent work has further demonstrated that microRNA (miRNA)-based regulation of PTEN can be modulated by the expression of competing endogenous RNA (ceRNA) targets. Several protein-coding RNAs that function as PTEN-regulating ceRNAs have now been experimentally validated [1]. However, the role of non-coding ceRNAs of PTEN has not been explored so far. In this work we present a machine learning approach for a large-scale identification of the non-coding ceRNA network of PTEN. We train a binary classifier on biologically relevant features extracted from the predicted target sites of PTEN-associated miRNAs.

Conclusions and Future Research

The supervised learning approach that we have proposed for identifying the PTEN non-coding ceRNA network suggests that the number of target sites of PTEN-associated miRNAs on the candidate ceRNA, along with the score that measures the quality of the target-site binding, are reasonable predictors of whether the candidate is a ceRNA of PTEN or not. The relatively low accuracy of our method is due to the small size of the training dataset and due to the fact that the negative examples are synthesized. We hope to remedy this issue in the future by considering more positive examples of experimentally validated ceRNAs of PTEN and using known non-ceRNAs as negative examples. This in turn will allow us investigate other predictors, such as the expression levels of the candidate ceRNAs in normal and tumor cells. The model, once trained on a more comprehensive training dataset, can be run on the entire genome to identify potential coding and non-coding ceRNAs of PTEN, which can then be experimentally validated. We would also like to explore other supervised learning algorithms, such as support vector machines.

Methods

We used a logistic regression model—a supervised learning algorithm—to distinguish PTEN-regulating ceRNAs from ones that are not. The training data for the binary classifier comprised of experimentally validated PTEN-regulating ceRNAs (see Table 1) as positive examples, and their random shuffles as negative examples. The 3'UTR sequences of the ceRNAs were shuffled using three different schemes, yielding three different training datasets: Γ_{ae} in which the ceRNA sequences where shuffled 100 times using the Altschul-Erickson dinucleotide shuffle algorithm [2]; $\Gamma_{\text{semi-random}}$ in which the ceRNA sequences were randomly shuffled 100 times; and Γ_{random} in which the sequences were random, but having the same length as the ceRNA sequences.

PTEN ceRNAs	ABHD13, CCDC6, CNOT6L2, CTBP22, DCLK11, DKK1, HIAT11, HIF1A2, KLF63, LRCH11, NRAS, RB11, SERINC12, TAF51, TNKS2, VAPA2
PTEN miRNAs	hsa-mir-17, hsa-mir-20a, hsa-mir-93, hsa-mir-106a, hsa-mir-106b, hsa-mir-20b, hsa-mir-519a, hsa-mir-519d, hsa-mir-18a, hsa-mir-216a, hsa-mir-217, hsa-mir-21, hsa-mir-141, hsa-mir-221, hsa-mir-222, hsa-mir-302a, hsa-mir-19a

Table 1: Experimentally validated PTEN-regulating ceRNAs and PTEN-associated miRNAs [1].

The features for the model were the number of target sites of PTEN-associated miRNAs (see Table 1) on the candidate ceRNA and the associated score, both computed using miRmap [3].

Results

Figures 1 (a)-(c) show the positive (blue) and negative (red) examples in the three training datasets, Γ_{ae} (a), $\Gamma_{\text{semi-random}}$ (b), and Γ_{random} (c). The x and y axes respectively denote the number of PTEN-associated miRNA targets on the candidate ceRNA and the corresponding miRmap score. The figures also indicate the decision boundary, i.e., the line separating the positive examples from the negative ones, computed by the logistic regression model.

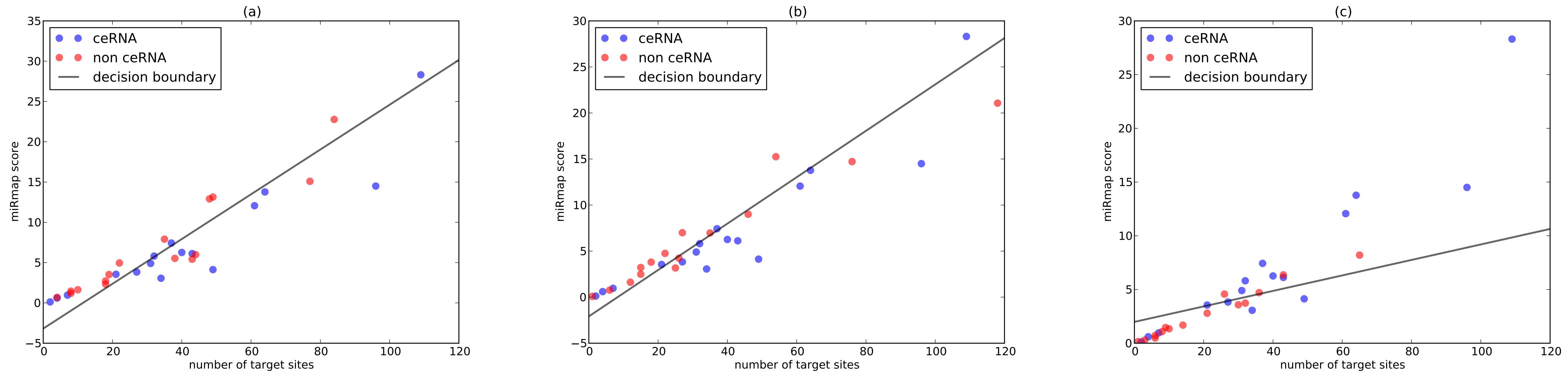

Figure 1: The three training datasets, Γ_{ae} (a), $\Gamma_{\text{semi-random}}$ (b), and Γ_{random} (c), along with the decision boundaries computed using a logistic regression model.

The in-sample accuracy of our model, i.e., accuracy within the training dataset, was measured using F-score, which is calculated as

$$\text{F-score} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of outcomes tested positive}} \quad \text{and} \quad \text{recall} = \frac{\text{number of true positives}}{\text{number of positives}}.$$

The out-of-sample accuracy, i.e, accuracy outside the training dataset, was calculated using leave-one-out-cross-validation (LOOC) as the fraction of samples that were correctly classified by the model. The F-score and LOOC accuracy values are shown in Table 2.

Training Dataset	F-score	LOOC Accuracy (%)
Γ_{ae}	0.62	65.62
$\Gamma_{\text{semi-random}}$	0.65	54.84
Γ_{random}	0.67	64.52

Table 2: The F-score and LOOC accuracy values for the three training datasets.

References

[1] Y. Tay, L. Kats, L. Salmena, D. Weiss, S.M. Tan, U. Ala, F. Karreth, L. Poliseno, P. Provero, F. Di Cunto, J Lieberman, I. Rigoutsos, and P. Pandolfi. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell*, 147(2):344–357, 2011.

[2] S.F. Altschul and B.W. Erickson. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Molecular Biology and Evolution*, 2(6):526–538, 1985.

[3] C.E. Vejnar and E.M. Zdobnov. miRmap: Comprehensive prediction of microRNA target repression strength. *Nucleic Acids Research*, 40(22):11673–11683, 2012.

Contact Information

http://www.cs.umb.edu/~swamir