Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites

Stefan Budach

12.11.2014

Seminar RNA Bioinformatics

Introduction

Betel *et al. Genome Biology* 2010, **11**:R90
http://genomebiology.com/2010/11/8/R90

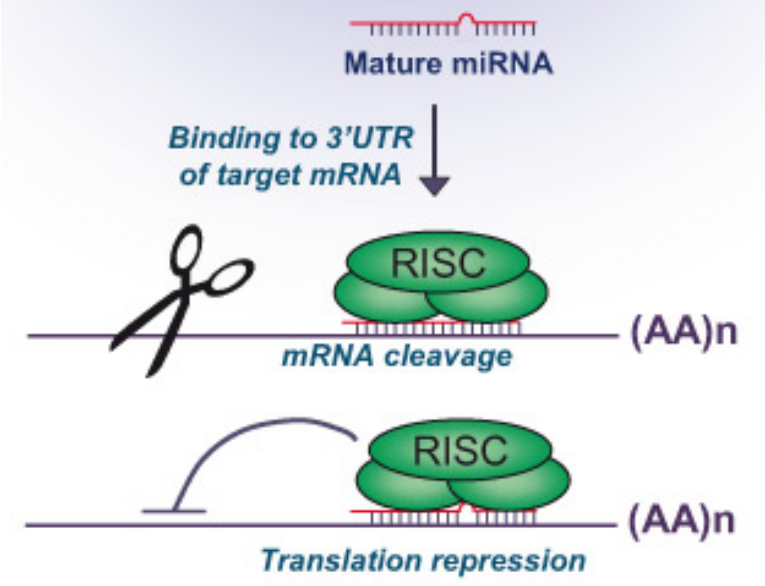
METHOD Open Access

Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites

Doron Betel[1], Anjali Koppal[2], Phaedra Agius[1], Chris Sander[1], Christina Leslie [1*]

Introduction

- miRNAs: small non-coding RNAs of length ~22 nt



Duroux-Richard et al., Swiss Med Wkly. 2011;141:w13175

- one miRNA can bind up to several hundreds mRNAs
- majority of mRNAs possess a conserved miRNA-binding site

→ miRNAs are an import part of regulation!

Introduction

- one challenge in the field of miRNAs: target site identification

- general problems:

  - experimental analysis is difficult
  - insufficient knowledge of miRNA biology
  - limited number of experimentally validated targets

- often by means of predictive computational methods

Introduction

- properties used by previous methods:
  - seed region (position 2 to 7 at the 5’ miRNA end)
  - most methods require perfect seed matching (= “canonical” site)
  - few methods allow G-U wobbles or other mismatches (= “non-canonical” site)
  - few methods: mRNA secondary structure
  - all methods: conservation filter
- problem:
  - lots of predicted sites, many false positives
  - conservation filter too strict

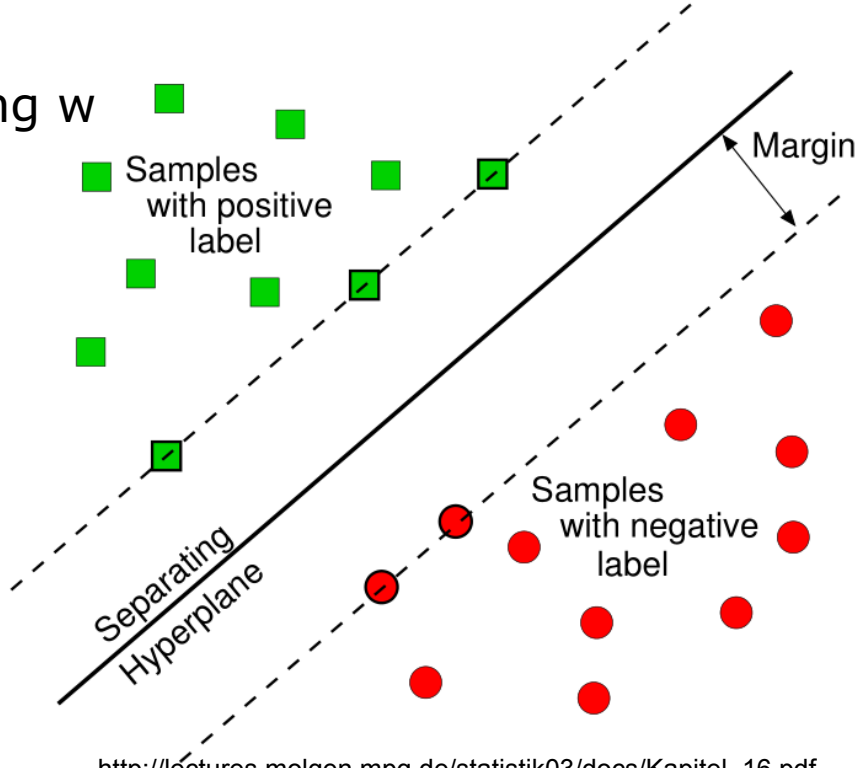
Introduction

- this paper: mirSVR
- support vector regression model for scoring and ranking the extent of down-regulation of given predicted target sites
- trained and tested on experimentally determined expression data
- prediction of poorly conserved and non-canonical sites without introducing a large number of spurious predictions

Support Vector Machine - SVM

- hyperplane: <w,x> + b = 0 margin: 1/||w||

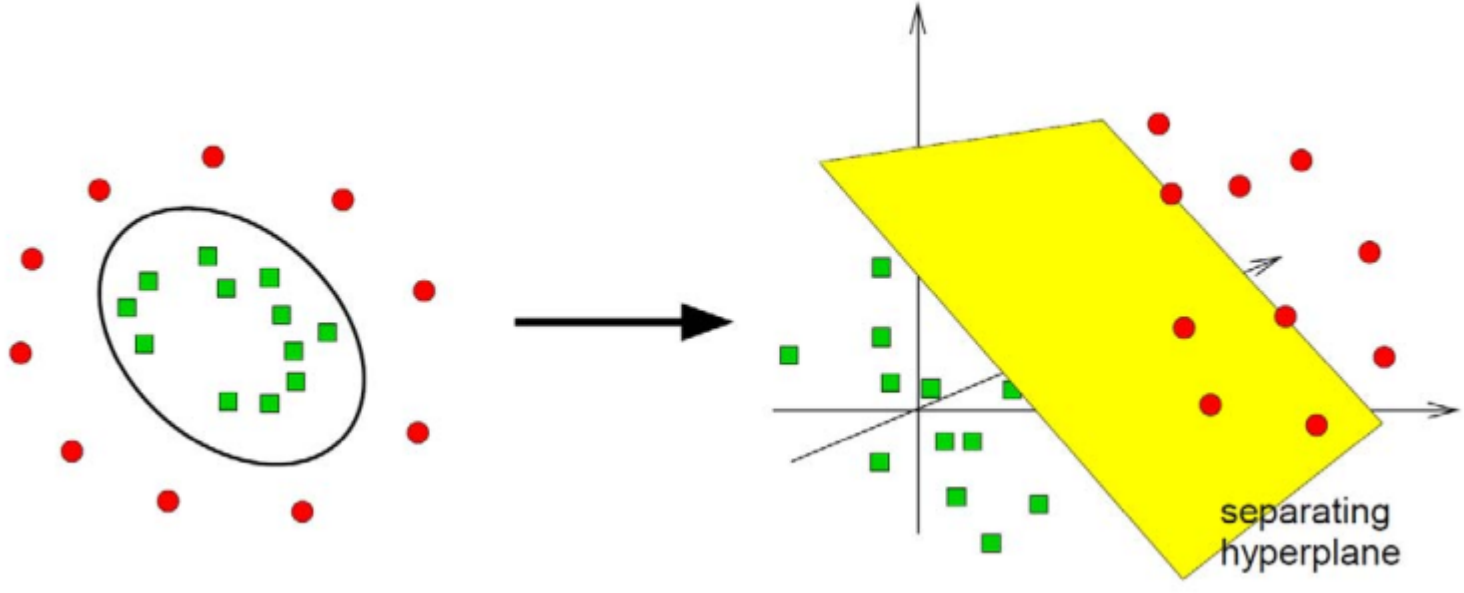
- main ideas:
  1) reduce data (sparsity)
  2) maximize margin by minimizing w
  3) linear separation in high-dimensional space (kernel trick)



http://lectures.molgen.mpg.de/statistik03/docs/Kapitel_16.pdf

Support Vector Machine - SVM

3) linear separation in high-dimensional space (kernel trick)



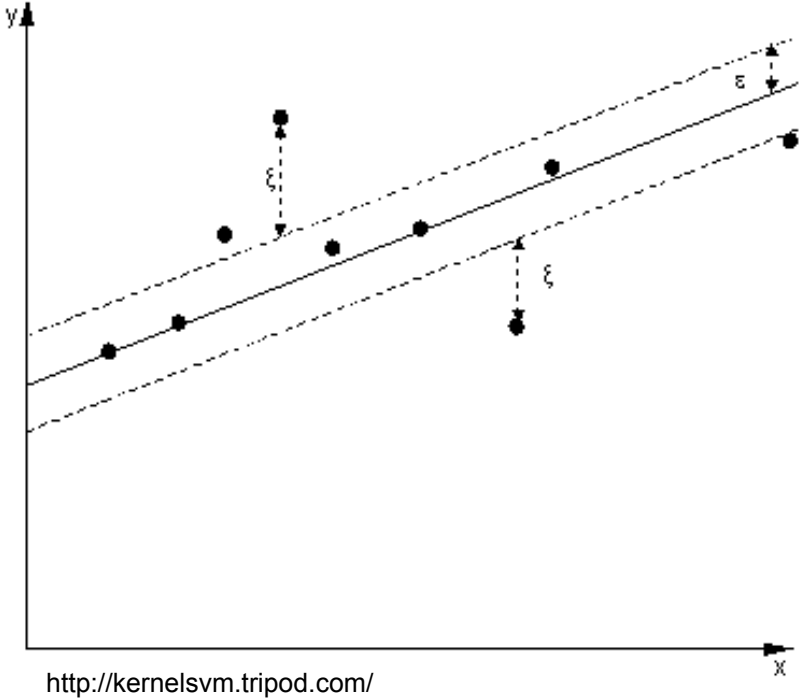
http://lectures.molgen.mpg.de/statistik03/docs/Kapitel_16.pdf

Support Vector Regression - SVR

- maximize number of points within an ε-tube, minimize error of points lying outside the tube

- linear regression using ε-insensitive loss function instead of e.g. ordinary least square:
  - errors less than ε are ignored
  - reduced data

- linear kernel (no mapping into high-dimensional space)



http://kernelsvm.tripod.com/

Data

- miRanda: alignment-based prediction of target sites
- mirSVR: ranking and scoring of target sites

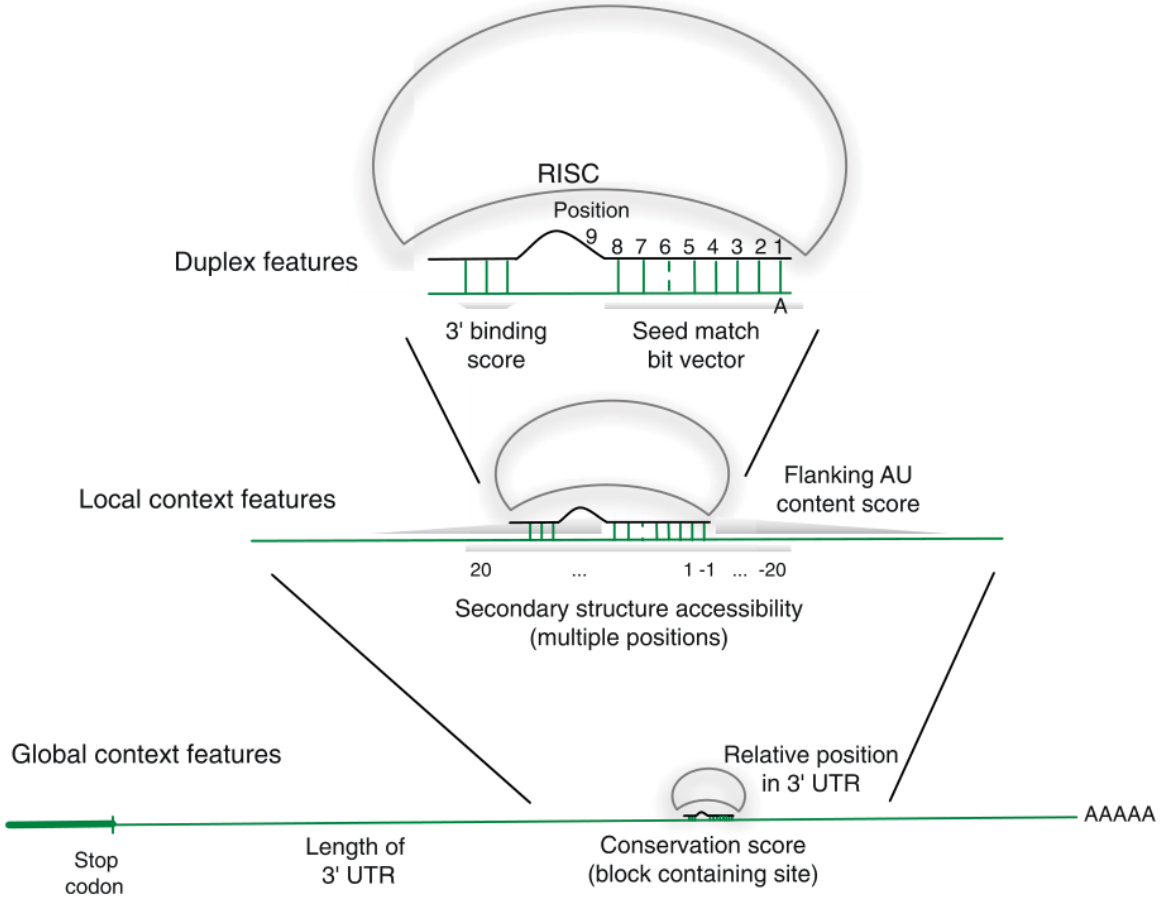
- expression data from transfection experiments:
  - measurement of mRNA expression levels
  - transfection of one microRNA into the cell (overexpression)
  - measurement of mRNA expression level 24h later

- prediction of log expression change using mirSVR

Data

- previously published data sets
- training data:
  - 9 mRNA expression array data sets
    using only genes with one target site
- test data:
  - 17 additional mRNA expression array data sets
  - 5 protein expression data sets
  - 3 inhibition experiments

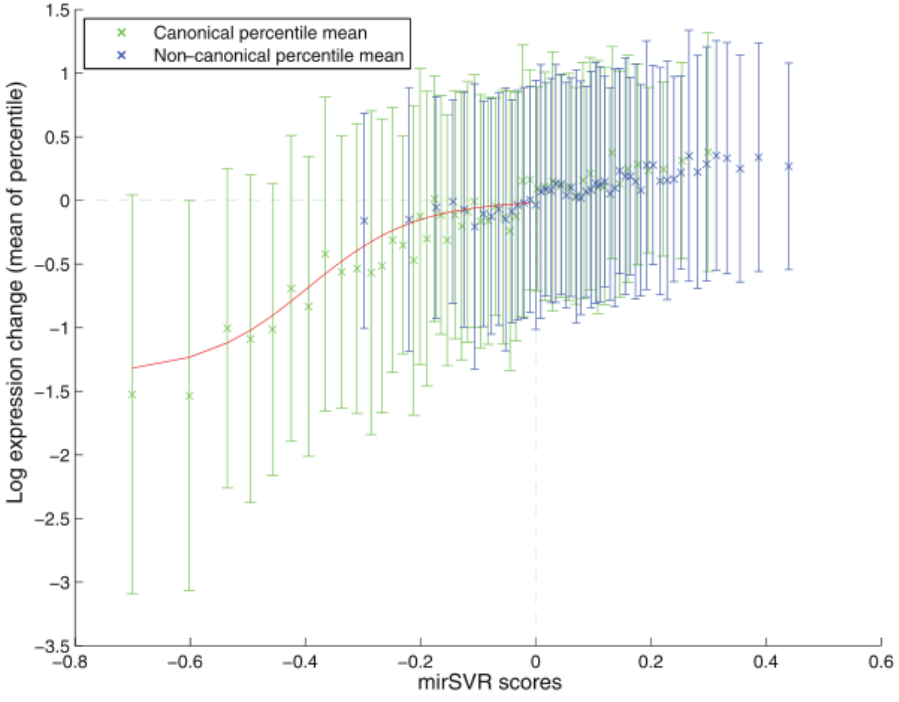
Model features

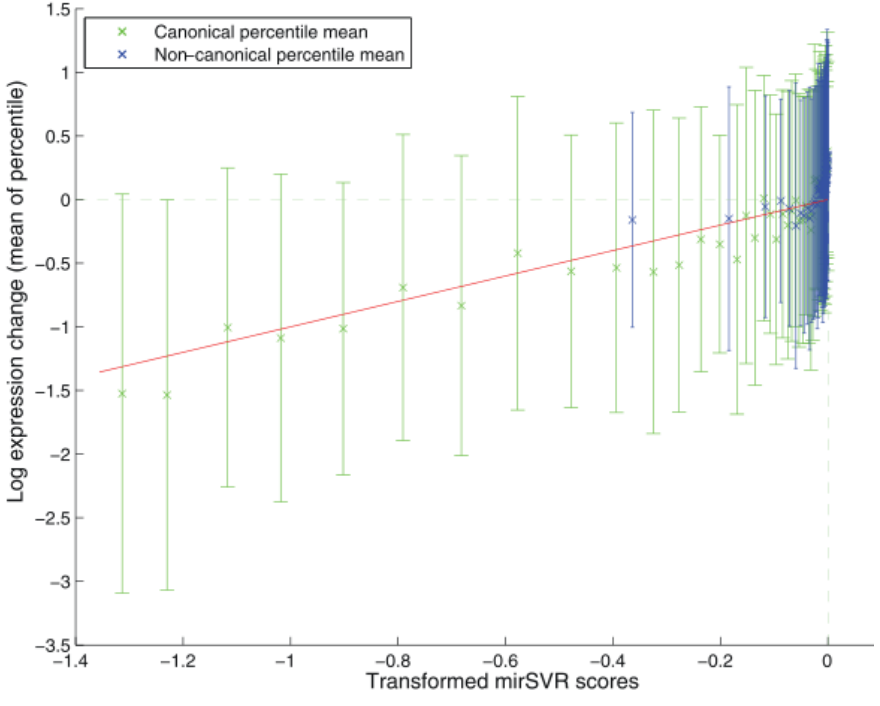


- features and log expression values are Z-score transformed (standardized) before training

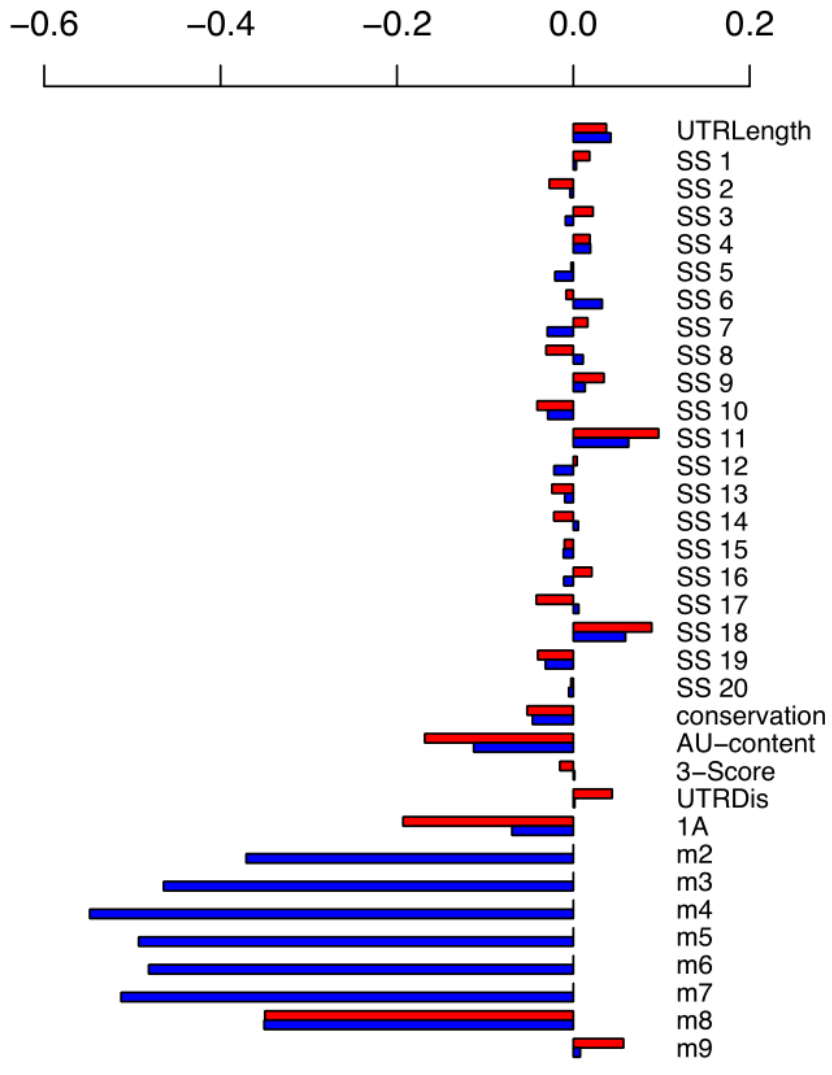
Score transformation

- score transformation with sigmoid transfer function to produce comparable values
- sum of scores for genes with multiple target sites

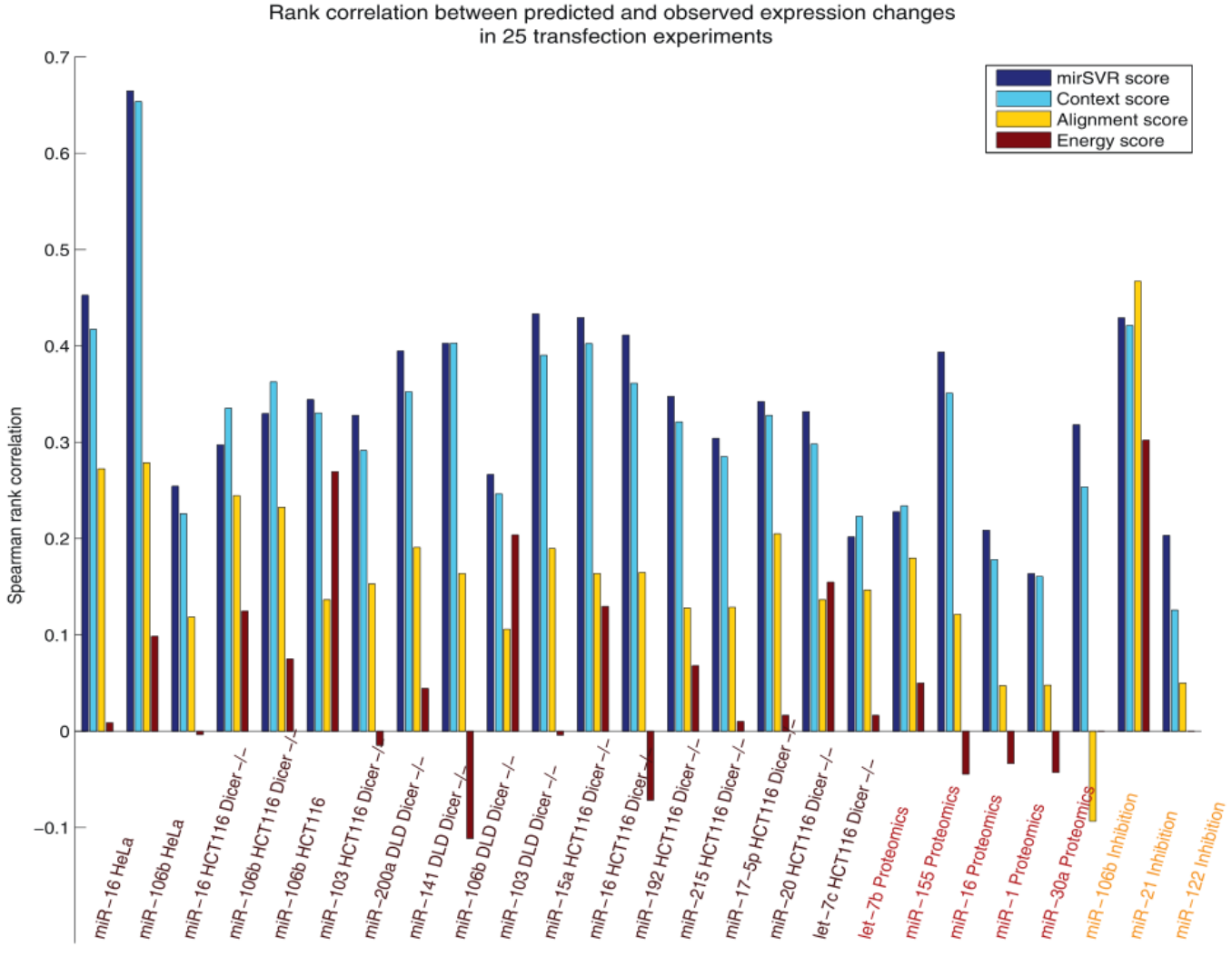




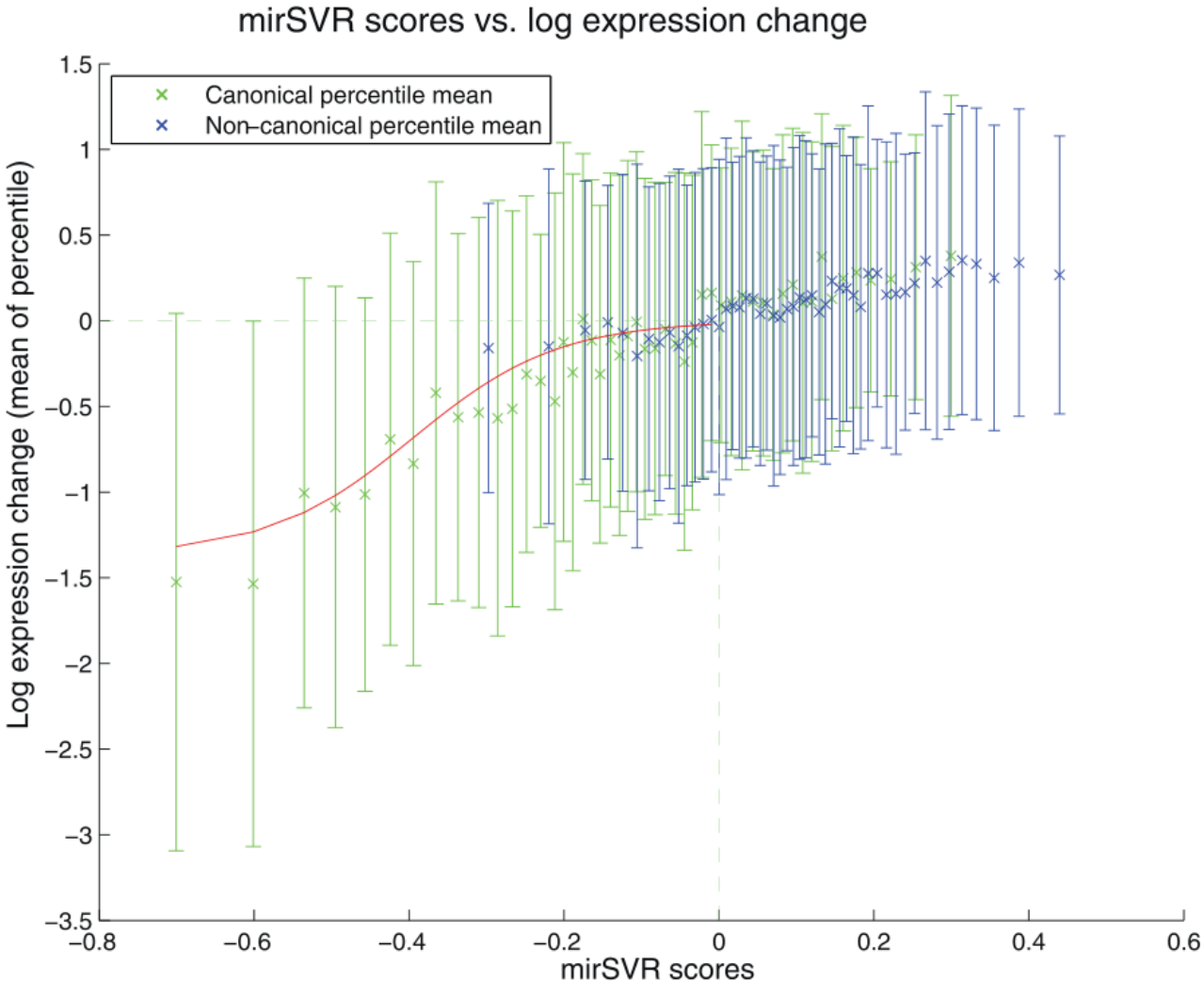
Regression coefficients



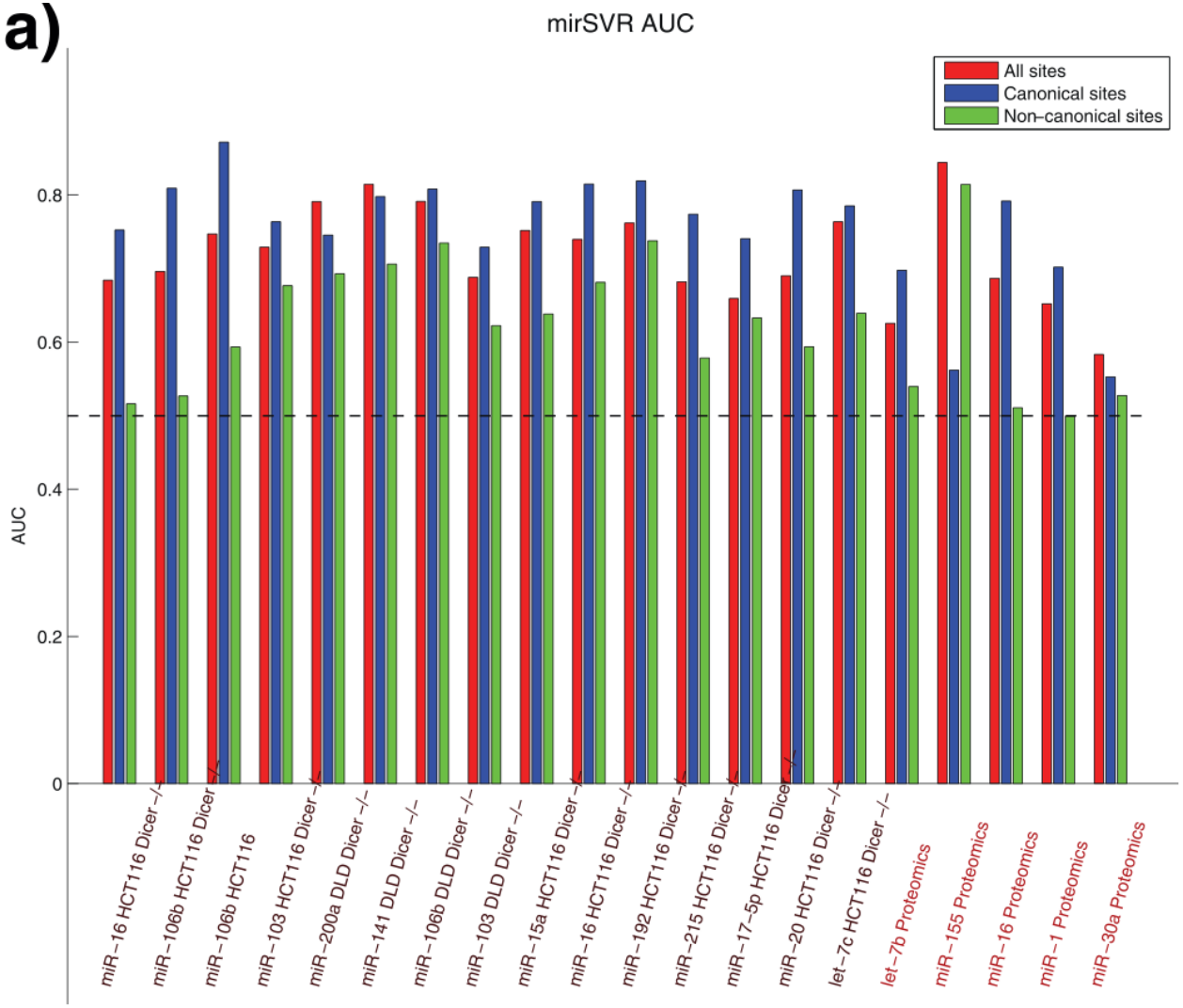
Spearman rank correlation



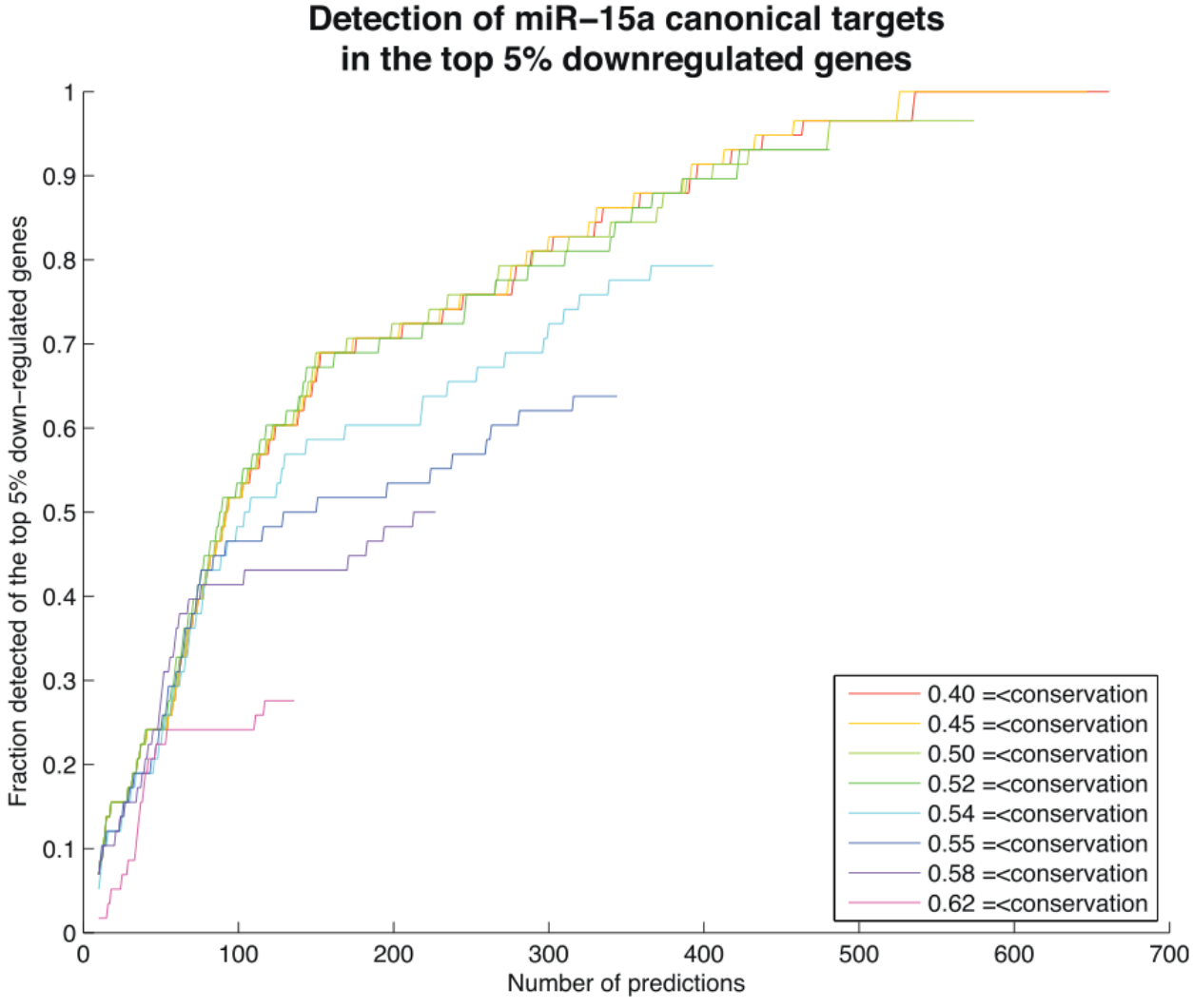
Non-canonical sites



Non-canonical sites



Conservation filter



Conclusions

- miRanda-mirSVR: competitive model due to
  - binary representation of seed features
  - training with robust SVR
  - score transformation to correlate with down-regulation
  - using conservation as a feature

- transfection experiments != physiological conditions
  - stronger regulatory effects
  - out-competing of other miRNAs
  - imbalance between miRNA and RISC concentrations

Future directions

- many target sites located in coding regions
- target specificity vary between different cell lines
- currently unknown non-sequence-specific features
- integrating RISC expression levels

Thank you for your attention!