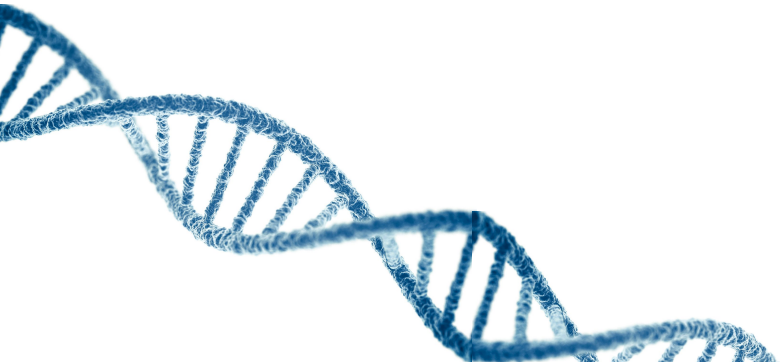
PREDICTING PROMOTER ACTIVITY FROM PROMOTER REGION SEQUENCE

Preksha Patel

Introduction

- Research Question: Developing a computational model to predicting promoter activity in yeast cells given the promoter region sequence

- Proposed Solution: Implement a linear regression model with regularization to predict promoter activity

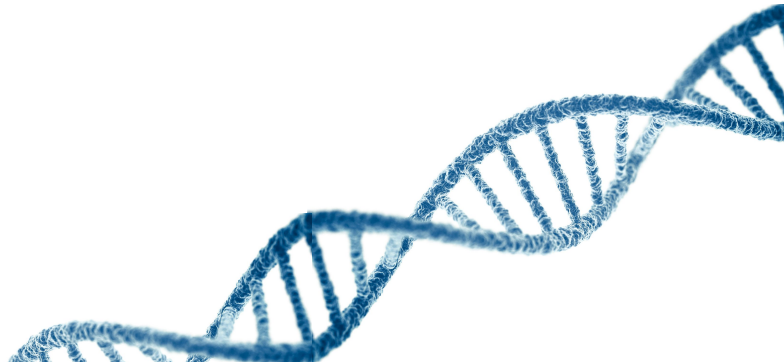


Methods

- Separate the training (90 sequences) and test data (53 sequences)

- Extract Features from the genetic sequence. Features include:
  - Frequency of k-mers (k=1-5)
  - Sum of the distance of the k-mers (k=1-5) from the gene
  - Number of occurrences of k-mers (k=1-5)
  - Number of occurrences of known motifs

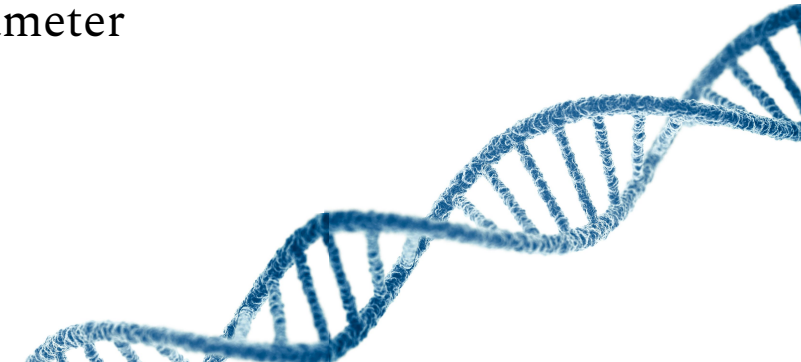
- Eliminate features with zero variance



Methods

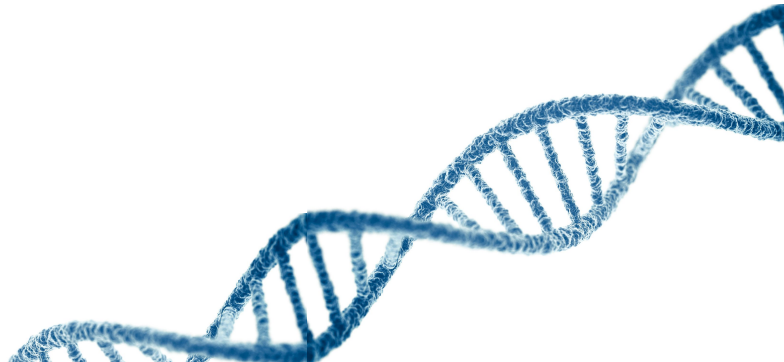
- Fit an unregularized linear regression model to establish baseline
  - R^2 regression score of -0.008 on test data

- Select a subset of features
  - Use recursive feature elimination
  - Train a linear regression model
  - Eliminate least significantly contributing coefficient
  - Treat number of features as a hyperparameter
  - Evaluate it using 5-fold cross-validation



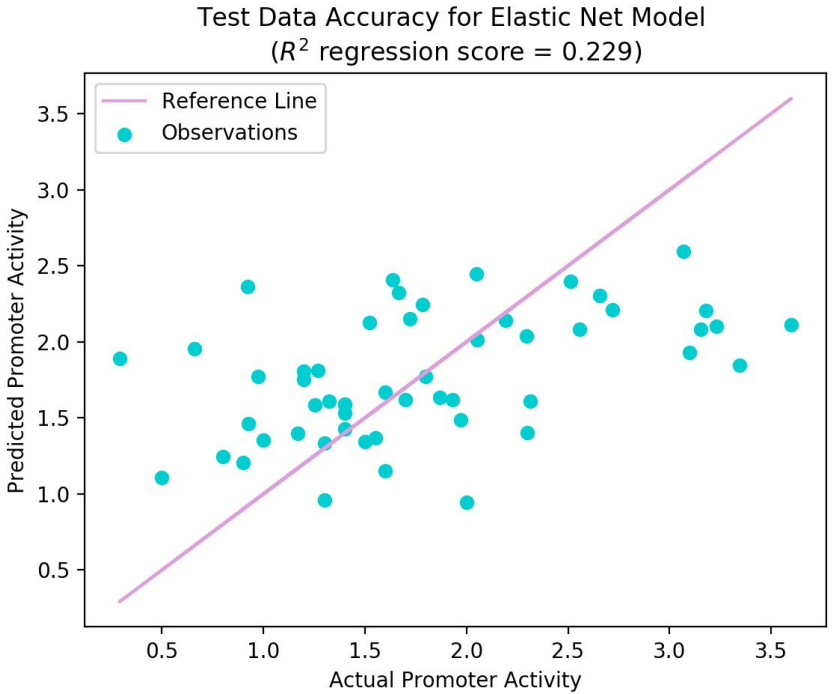
Methods

- Fit regularized linear regression models
  - Models include Ridge, Lasso and Elastic Net
  - Estimate hyperparameters using 5-fold cross-validation
  - Perform a grid search
  - Train the model on the entire training dataset
  - Evaluate performance using test data



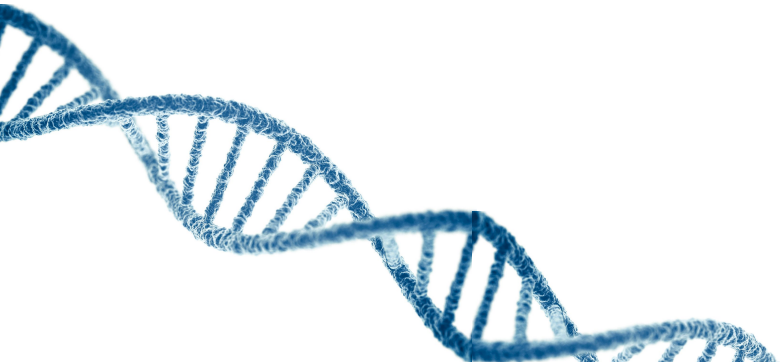
Results

- Best performing model was elastic net with no feature selection
- Performance on test data
  - R^2 regression score of 0.229
  - Spearman correlation of 0.491
- Predicted promoter activities were concentrated in a smaller range



Discussion

- Possible reasons for low R^2 regression score (=0.229)
  - Feature extraction resulted in a loss of essential information
  - Small size of training dataset
  - Relation is non-linear



THANK YOU!