PREDICTING PROMOTER ACTIVITY FROM PROMOTER REGION SEQUENCE

Preksha Patel

INTRODUCTION

The expression of genes is largely regulated by the DNA sequence upstream of it, known as the promoter region. Understanding the influence of the promoter region on gene activity is fundamental to our comprehension of natural variations in gene expression. To this extent, various models have been developed to predict the promoter activity from the promoter region DNA sequence. As a solution to the DREAM challenge, Siwo, et al. (2016) proposed a support vector machine model to predict promoter activities in yeast. Their model predicted promoter activities with a spearman correlation of 0.73 in natural sequences and 0.57 in laboratory mutated sequences. However, there is potential to improve the accuracy of the predictions.

The proposed hypothesis is that a linear regression model could be implemented to predict the promoter activity from the promoter sequence. To build a model, features were extracted from the sequence including frequency of k-mers (k=1-5), sum of position of k-mers, number of occurrences of k-mers, and number of occurrences of known motifs (Badis, et al., 2008). Feature selection and regularization (Ridge, Lasso, and Elastic Net) were implemented to avoid overfitting. In addition, hyperparameters selected for the regularization models were optimized with a grid search using 5-fold cross-validation. Using the optimized hyperparameters, the model was trained on the training dataset. The accuracy of prediction on the test data indicated that an elastic net model without feature selection performed the best (R^2 regression score of 0.229). In the future, the accuracy of the model can be improved by using a larger sample size to train the model, by selecting more relevant features, and by exploring non-linear models.

METHODS

The DREAM challenge data comprises of 90 DNA promoter region sequences in the training set and 53 sequences in the test set. The challenge defined the promoter region as the sequence starting immediately upstream of the ribosomal protein-coding region and extending until the previous gene sequence, or if shorter, until 1200 base pairs. Each promoter sequence was inserted into the same fixed position upstream of a yellow fluorescent protein (YFP) gene in the yeast genome. The measured YFP expressions per cell per second were provided as a metric for the activity of the promoter region. Since all strains of yeast had been grown in the same environment, the variation in activity was solely due to the variations in the promoter region sequences. Consequently, it was assumed that certain features of the DNA sequence would be correlated with the promoter region activity. Features were extracted from the DNA sequence including the frequency of k-mers (k=1-5) in the promoter region, the sum of the positions of the k-mers, the number of occurrences of k-mers, and the number of occurrences of known motifs. Once the set of features was defined, all the features with a zero variance were eliminated.

After determining the complete set of features, the performance of linear regression models with different methods of regularization was evaluated. The models were trained on the training data (90 promoter region sequences), and their accuracy was determined on the basis of their performance on the test data (53 promoter region sequences). To establish a baseline, an unregularized linear regression model was initially implemented. The model gave an R^2 regression score of -0.008 on the test data which was much lower than its R^2 regression score of

1.0 on the training data. These preliminary results indicated overfitting. To overcome the problem of overfitting, feature selection and regularization were implemented.

Feature selection helps reduce the number of features used to train the model, which could potentially solve the problem of overfitting. The feature selection method implemented was recursive feature elimination. Starting with all the features, a linear regression model was trained and then features with the least significantly contributing coefficient were eliminated. This was repeated until a set number of features were obtained which were used to train the model. The number of features selected was treated as a hyperparameter and was optimized using five-fold cross-validation. Once the optimal number of features was determined, that subset of features was used to train the model.

Regularization was implemented in the form of ridge regression, lasso regression, and elastic net models. For each of these, the hyperparameters were evaluated through five-fold cross-validation on the training data. A grid search strategy was utilized to select the optimal value for the hyperparameters. The regularization parameters were optimized over an exponential grid. After evaluating the hyperparameters, the model was trained on the complete training set using the optimal values of the hyperparameters. The accuracy of the models was estimated by predicting the promoter region activity for the test data and computing an R^2 regression score for the test data.

RESULTS

The baseline unregularized linear regression without feature selection gave an R^2 regression score of -0.008 on the test data and a score of 1.0 on the training data. The score on the training data as compared to the score on the test data was highly indicative of overfitting, so various regularized models and feature selection methods were implemented to prevent overfitting. After implementing linear regression, ridge regression, lasso regression and elastic net, with and without feature selection, it was observed that the elastic net model without feature selection gave the best prediction results on the test data (R^2 regression score of 0.229). The performance of elastic net and lasso regression was similar, while ridge regression and unregularized linear regression performed poorly when compared to elastic net and lasso.

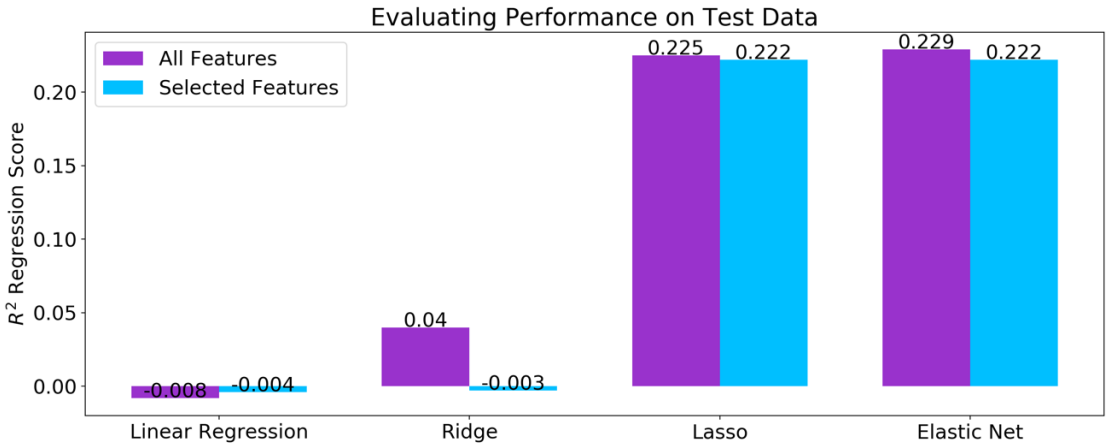

Fig 1. Performance of different models on test data as measured by the R^2 regression score.

Additionally, while feature selection increased the R^2 regression score for linear regression, it did not improve the performance of any other model (Fig 1). Without feature selection, both lasso and elastic net had 28 features, while ridge and unregularized linear regression had 2698 features. Feature selection reduced the total number of features from 2698 to 311. With feature selection, lasso had 25 features and elastic net had 26 features while ridge and linear regression had 311 features. In general, models that performed better had significantly fewer contributing features.

Though the elastic net gave a better accuracy compared to the baseline model, the R^2 regression score was still not that high, indicating a high prediction error. The model did not effectively capture the factors predicting the promoter activity. Among the predicted results, the model performs better when predicting the promoter activity in the range of 1.0 to 3.0 (Fig 2). However, it performs poorly on the rest of the data. This could be due to the fact that most of the training data was concentrated in that region (87%). It was also observed that the predicted promoter activities by the model were in the range 0.94 to 2.59 while the actual promoter activities of the test data ranged from 0.29 to 3.59.

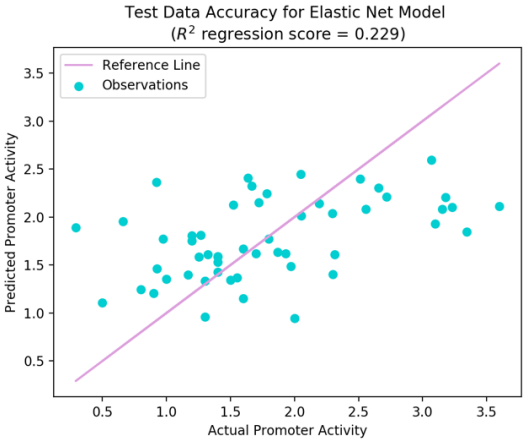

Fig 2. Promoter activity predicted on the test data from an elastic net model without feature selection

DISCUSSION

The elastic net model implemented in this project had a low level of accuracy in predicting actual promoter activities from the promoter sequence. This model would not be a good choice for predicting promoter activity from promoter region sequences due to the low R^2 regression score on the test data. The low R^2 regression score could have been due to multiple reasons. It is possible that feature extraction from the promoter region sequences resulted in a loss of essential information needed to predict promoter activity. That is, the features were not representative of the data. Moreover, it is possible that the model was overfitting the data and that the problem was not completely solved by regularization or feature selection. The small sample size of the training data set could also have been insufficient to learn the predictive features in the promoter sequence. All these could have been potential issues with the current linear models, but if the relationship between the defined features and promoter activity is not linear, then the assumption of linearity would not hold, and a linear model would not fit well.

In the future, the model could be improved by working on these aspects. Non-linear models could be implemented to predict the promoter region activity. In addition, other features that are highly correlated with promoter activity could be extracted from the sequence. While these might help improve the model, the ideal solution would be to train a model on a larger sample size of data. Altogether, there is a lot of potential to improve this model and obtain a better performance on the training data.

REFERENCES

1. Siwo, G., Rider, A., Tan, A., Pinapati, R., Emrich, S., Chawla, N., & Ferdig, M. (2016). Prediction of fine-tuned promoter activity from DNA sequence. *F1000Research*, *5*, 158. https://doi.org/10.12688/f1000research.7485.1
2. Badis, Gwenael et al. “A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters.” *Molecular cell* vol. 32,6 (2008): 878-87. doi:10.1016/j.molcel.2008.11.020
3. DREAM Gene Expression prediction Challenge. (n.d.). Retrieved from https://www.synapse.org/#!Synapse:syn2820426/wiki/