**­­I529 Group Project - Gene expression prediction**

Will Shoemaker, Xiangyu Yao and Meng Wu

**Introduction**

Gene expression is largely influenced by the promoter region upstream the gene. In this study, we investigated how different characteristics of promoters can be used to predict gene transcription activity. In particularly, we pulled out the information of promoters from the training dataset, such as promoter length, motif copy numbers, motif position, and GC contents of promoters. We then performed cross-validation using a linear model on predicted vs. observed expression values of different models. Our data suggested that one predicted motif position and GC-content are the most important promoter characters to predict expression level of target genes.

**Methods**

*Motif Prediction -* Based on our knowledge of common promoters, clusters of motifs lie around upstream of transcription start sites (TSSs), which provide transcription biding binding sites (TFBS) for corresponding transcription factors. A single gene may use alternative TSSs and corresponding motifs in different tissues or under different environment, explaining the basic mechanism of gene regulation. For our final project, to identify motif clusters in promoters is the first step to build model predicting promoter activity score. We performed *de novo* searching on TFBS motifs using an integrated motif-predicting tool MEME. The program can return the highest scored position-specific scoring matrix (PSSM) of predicted motifs by using EM algorithm. Since we used 10-fold cross-validation to measure our model in the following step, we randomly split our 90 promoter raw data set into 10 groups and randomly select 9 of them as training set and the remaining one as test set. We set the parameters in the MEME to finds at most 5 motifs by allows any number of repetitions of motifs, motif width in range of 6~15 bps. We then searched our predicted motifs in the motif dataset TOMTOM.

*Regression analysis -* Motif count data was cleaned from the MEME output file and merged with a data frame containing sequence GC-content and the log10 transformation of sequence length as well as the number of motifs in each sequence, the mean relative position of each motif, and the p-value of each motif. Because we are using a relatively small dataset with relatively few features, ordinary least squares linear regression was used. All analyses were conducted in Python 2.7.11 and Scikit-learn 0.17.1. Exploratory data analysis was done using scatterplots. Recursive feature elimination with cross-validation using linear regression was done to find the optimal number of features to use. Linear regression models were constructed for the entire dataset and on the optimal features using 10-fold cross validation. Residual plots were made for the predicted expression based on the entire dataset. Observed versus predicted plots were made for the entire dataset. Scoring of each model was done as assigned, using an overall score metric. The metric used the p-values for Pearson’s correlation coefficient, Spearman’s rank correlation coefficient, the Chi-squre goodness-of-fit test, and the Wilcoxon signed-rank test using SciPy ‎0.17.0. The following equation was used for the overall score metric for all four p-values:

.

All code and data is available at the following GitHub repo:

https://github.com/wrshoemaker/I529-Group-Projects/tree/master/FinalProject

**Results**

Based on the 5 predicted motifs (Fig. 1), only the first one seems to be a real motif and have a significant hit with known motifs in TOMTOM dataset. The other four motifs are either poly-A like sequences or with high E-values. But we used all of those 5 predicted motifs for model 1 analysis (Fig. 1). Recursive feature elimination with cross-validation suggested that only GC-content and the mean relative position of the fifth motif should be selected. These two features were analyzed as model 2. There is a high amount of variation in the residuals for model 1 (Fig. 2) as well as model 2 (not pictured). Our results suggest very modest predictive power in our linear regression (Fig. 3), though our results are significant (Table 1). The overall score is on par with the leaderboard for the [Dream6 Gene Expression Prediction Challenge](https://www.synapse.org/#!Synapse:syn2820426/wiki/71013), as our score for model 1 currently puts us in fourth place.

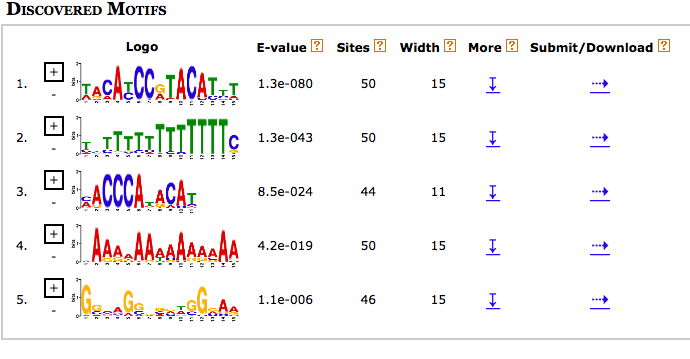


Figure 1. Top 5 predicted Motifs

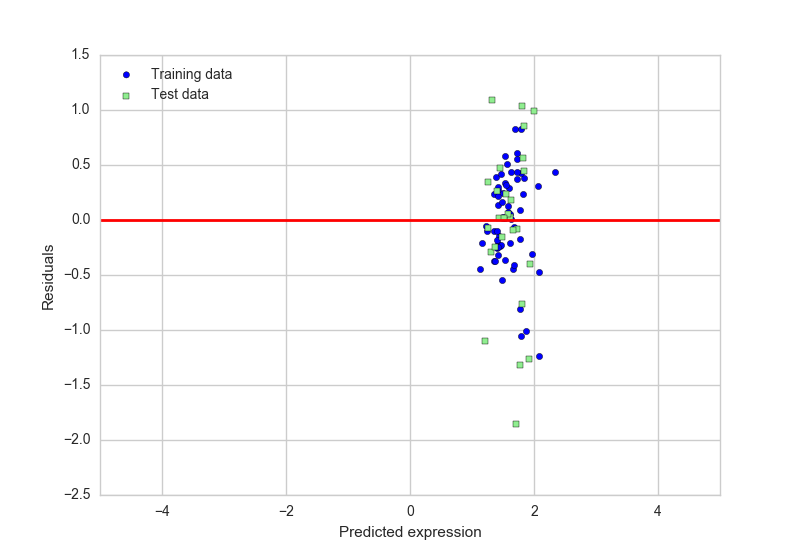


Figure 2: A plot of predicted expression versus residuals for all features (model 1).

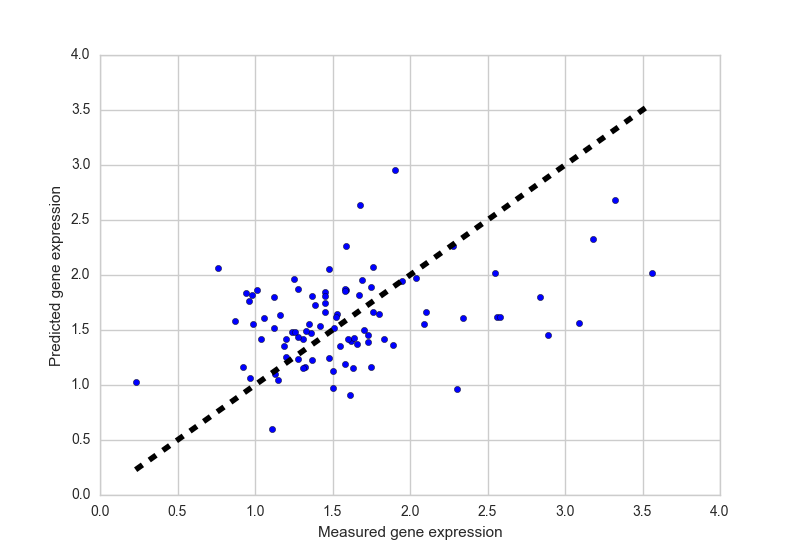


Fig. 3: Measured versus predicted gene expression for all features in the data set (model 1).



Table 1: The overall score, coefficient, and p-value for each statistical test or analysis for model 1 and 2.

**Discussion**

Our motif prediction may not be accurate, even though we have tried different parameters in MEME. The reason may be that we cannot have a reasonable motif that predicts gene expression. This could either be due to issues with MEME or due to a lack of data. However, our results are significant and suggest that, at some level, gene expression can be predicted using data from motif finding and sequence features.

**Statement of authorship:** WS, MW, and XY conceived and designed the study and

reviewed and wrote the paper. MW and XY identified the motifs. WS formatted the data and conducted the regression analysis.