Nicolas Lynn - 946747680
Michal Eck - 026511378
Lotem Dallal - 204168348

# Gene Repression Modeling

**Analysis 1:**

We began our analysis by first selecting and separating our data. We ignored all miRNA-Gene combinations that had more than a single binding site combined (between UTR3 and ORF). Initially, included analysis of the UTR5' sequence. We realized after that perhaps it wasn't as good of a choice because there were about 10% as many legitimate binding sites in the UTR5 as either UTR3 or ORF. The smaller the number of samples, the easier it was to overfit the data.

Our first round of analysis required that we look at several single features alone. A list of all the features we looked at is below. The following features were generated for each of: **UTR5**', **ORF**, and **UTR3**'

1. **Folding energies** of a 78 nucleotide window around binding indices (70 + 7 + 1, 70 is the window 7 is the seed and one is added so the total length is divided by 3).
2. **Length of the miRNA** and average repression across all genes for that miRNA. This was an additional feature we suggested.
3. **Length of the gene** segment (UTR5', ORF, UTR3') and average repression across all miRNAs for that gene. This was an additional feature we suggested.
4. **Conservation** of binding window, the whole sequence, and a ratio. The two latter were our own additional features.
5. **Distance to terminus** of binding index (5' terminus, to 3' terminus, and to either), and a ratio compared to the whole segment length.
6. **CAI** of respective ORF for each valid miRNA-gene combination.
7. **tAI** of respective ORF for each valid miRNA-gene combination.
8. **GC content** of binding window, the whole sequence, a ratio of the two, as well as of the UTR5, ORF, and UTR3.
9. **Codon count percentages** for the whole sequence and for the binding window (this resulted in 64 additional features for each). This was an additional feature we believed would be beneficial.

**Total Features: 150**

**ORF Total Binding Sites Found: 7688**

**UTR3 Total Binding Sites Found: 7549**

How we created these features can be found in the MATLAB code. We determined each feature's value by finding the Pearson and Spearman Correlation between the observed repression values and the predicted repression values for each individual feature. We found that there was very little correlation (between 0% and 4%) for each individual feature. Regardless, from this initial analysis, we were able to conclude that some features had a more significant relationship than others. In particular, we found that the length of each miRNA and their average repression values across all genes had the highest relationship. The correlation of the values observed and predicted with this feature was 16.33%. However,

it should be noted that the number of unique lengths of the miRNAs was only 4. The relationship between the length of the sequence and the average repression values across their miRNA binding combinations was also relatively significant, with the predicted-observed repression correlation at 11.57% for ORF sequences. THIS WAS THE BEST RELATIONSHIP OBSERVED AS THERE WAS LITTLE LIKELIHOOD OF OVERFITTING (~4000 samples and only one feature). Besides these 2 features, the range of correlation for all others was between 0.9%-4.5%. Alone, none of these features can be considered to be well correlated. However, this is to be expected as gene repression is a highly complicated process that is dependent on numerous factors, many of which may not even be known, let alone described in this data. The exact correlations of each feature can be observed simply by running the the main.m MATLAB script.

**Analysis 2:**
Next we wanted to try putting some of the data together. We used three regression types: lasso, stepwise regression, and simple regression.

We will break our findings by grouping of segment: UTR5', ORF, UTR3'

**** Ignore ** UTR 5':** We were initially surprised at the results we found from the UTR5' analysis. It was an additional — not recommended — consideration that ended up with good results, though likely very overfitted. There were only around 500 found binding indices, approximately 10% of the quantity found in the others.

**Correlations**

| **UTR5** | **Lasso** | **Stepwise Regression** | **Simple Regression** |
|---|---|---|---|
| **Pearson** | 0.22 (22.575) | 0.33 (32.95%) | 0.23 (22.67%) |
| **Spearman** | 0.16 (16.82%) | 0.24 (24.86%) | 0.17 (16.89%) |

Selected features:

I. CAI and TAI of ORFs
II. Conservation of Binding Window
III. Folding Energy of Binding Window
IV. Distance to Nearest Terminus
V. GC Content of UTR5, ORF, and UTR3
VI. Length of UTR5, ORF, and UTR3

**Correlations**

| **UTR5** | **Lasso** | **Stepwise Regression** | **Simple Regression** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| **Pearson** | 0.173 (17.28) | 0.18 (17.9%) | 0.17 (17.25%) |
| **Spearman** | 0.13 (13.00%) | 0.13 (12.82%) | 0.13 (13.09%) |

We will not discuss the results of UTR5 since it is irrelevant.

**ORF:** We were slightly disappointed at the fact that we could not obtain as good results with the ORF region, which we initially expected we would. We first used all 150 features and ran them through all three regression types. The lasso regression used a cross validation of 10 and an alpha parameter of 0.5.

**Correlations**

| **ORF** | **Lasso** | **Stepwise Regression** | **Simple Regression** |
|---|---|---|---|
| **Pearson** | **0.14 (14.39%)** | 0.0877 (8.77%) | 0.071 (7.1%) |
| **Spearman** | 0.14 (14.03%) | 0.093 (9.28%) | 0.064 (6.4%) |

Clearly, the lasso results were best. Next, we tried running the regression only on the recommended features. We experienced a serious decrease in correlation.

Selected features:

I. CAI and tAI of ORFs
II. Conservation of Binding Window
III. Folding Energy of Binding Window
IV. Distance to Nearest Terminus
V. GC Content of UTR5, ORF, and UTR3
VI. Length of UTR5, ORF, and UTR3

**Correlations**

| **ORF** | **Lasso** | **Stepwise Regression** | **Simple Regression** |
|---|---|---|---|
| **Pearson** | 0.058 (5.8%) | 0.073 (7.3%) | 0.0556 (5.56%) |
| **Spearman** | 0.0482 (4.8%) | 0.0576 (5.76%) | 0.0481 (4.81%) |

The fact that there is such a dramatic impact on the cross validation values (even when we simply removed other features) shows that this model is rather fragile and not heavily reliant on a single feature. This makes us believe that this particular model would not be a great predictor for other sets of data. However, because there are a total of approximately 21 predictors (clustering the codon counts which

account for 64 features each into 2 categories), we decided to use all the predictors with the lasso regression model.

**UTR 3':** We did not get very different results with the UTR3' region.

**Correlations**

| UTR3 | Lasso | Stepwise Regression | Simple Regression |
|---|---|---|---|
| **Pearson** | **0.16 (15.74%)** | 0.12 (12%) | 0.0944 (9.4%) |
| **Spearman** | 0.13 (13.18%) | 0.090 (9.02%) | 0.071 (7.1%) |

Ocne again, we used the recommended features to determine any correlation with a handful of standard features. The drop in correlation was very similar to the drop seen in the ORF segment.

Selected features:

I. CAI and TAI of ORFs
II. Conservation of Binding Window
III. Folding Energy of Binding Window
IV. Distance to Nearest Terminus
V. GC Content of UTR5, ORF, and UTR3
VI. Length of UTR5, ORF, and UTR3

**New Correlations**

| UTR3 | Lasso | Stepwise Regression | Simple Regression |
|---|---|---|---|
| **Pearson** | 0.0739 (7.39%) | 0.083 (8.3%) | 0.0731(7.31%) |
| **Spearman** | 0.0629 (6.29%) | 0.0602 (6.02%) | 0.0613 (6.13%) |

The results from our trial confirm that the process of miRNA binding and its effect on repression are not well understood and not well described by a combination of measures such as tAI, conservation, GC content, or other described here, let alone on a single one. The best result of this model is likely not very useful for other data as we believe it is overfit to this data. If we were able to use it (we are not allowed to in order to be within the bounds of the assignment), we would use the average gene length versus average repression, disregarding miRNAs completely as we saw a relatively strong relationship in correlation values (0.12) with only a single feature, which is likely not overfit.

Leaving aside the complicated results of this exploration, there were two single features that did not apply on a per miRNA-gene combination, rather on miRNA length and sequence length on average. Using a simple regression, we found that the length of the sequence containing a binding site has a relatively large (compared to the other values found here) correlation with the observed repression in the ORF.

| **Simple Correlation** | **UTR'** | **ORF** | **UTR3'** |
|---|---|---|---|
| **Pearson** | -0.0177 (1.77%) | -0.1157 (11.57%) | 0.053 (5.3%) |

We performed a similar regression on miRNA length and average repression across all genes and found a correlation in predicted results of 16.33%, yet there was a range of 4 possible values of miRNA length in this set of data which does not allow us to make any statements into the miRNA length and repression relationship.

After all, we are able to conclude that miRNA binding site and repression are not easy to relate. There is certainly a complex and not fully understood mechanism governing gene repression.

The reason we selected this model is because we believe that it is not more overfit than the other models we generated and because there were no single features that consistently provided stronger predictions and held higher weights in terms of predictions.

**Validation**

For the repression predictions in the validation set, we will use:

**FINAL MODEL: LASSO WITH 150 FEATURES, CV 10, ALPHA 0.5**

FOUND: 103 ORF binding sites, 193 UTR3' binding sites.

Find the final predictions under >> **validation_predictions/predictions.mat**

The format of the final predictions are one 1 x ~5000 arrays containing the prediction values for each segment (ORF, UTR3') that had a binding site (one a single binding site between two sites). Because there were 11 combinations where there was only a binding site in the UTR5' (which we ignored), 103 combinations where there was only one binding site in the ORF, and 193 combinations where there was only one binding site in the UTR3', there are a total of 296 predictions in total.