The International Young Researchers' Conference
DOI: 10.34614/iyrc2021f43

Investigating the Sequence Elements that Affect the Translation Efficiency of the Fungal Pathogen Histoplasma Capsulatum

Annika Viswesh
Palo Alto Senior High School, United States
annikaviswesh@gmail.com

Abstract

Histoplasmosis disease is caused by the dimorphic switch of the Histoplasma capsulatum fungus. Predicting the Translational Efficiency (TE) for Histoplasma capsulatum will lead to techniques that can regulate its protein production and thereby help in the treatment of Histoplasmosis. However, what sequence elements in the mRNA determine TE in Histoplasma is not well understood. The 5' Untranslated region (UTR) of 4981 genes common to 4 strains of Histoplasma were explored to identify the correlation between the longest 5 Upstream Open Reading Frame (uORFs) with start codon ATG, length of the 5' UTR, the energy of constrained secondary RNA structure, CG-to-ATG ratio and TE, using Wilcoxon tests, normal distribution plots, and Area under the receiver operating characteristics (ROC) curve. Subsequently, using all these sequence elements as features, four computational models were developed using different machine learning algorithms to predict TE. The results demonstrate that the maximum length of uORF with start codon ATG and the CG-to-ATG ratio have the best correlation to TE with the highest Area Under the Curve (AUC) amongst all sequence elements at 0.74 and 0.79, respectively. Also, computational model created using Random Forest outperformed other models to best predict TE with an AUC of 0.85. This research helped identify a set of sequence elements that affect TE in Histoplasma capsulatum and also showed that computational models can be created for predicting the TE of Histoplasma.

Keywords: Histoplasma capsulatum, Histoplasmosis, machine learning, sequence elements, translation efficiency

Introduction

Histoplasma capsulatum is a fungus that has a yeast form and a mold form (Gilmore et al. 2015; Inglis et al. 2013). They reside in soils. They are commonly found in the Ohio and Mississippi River Valleys, as well as in other parts of the world (Bahr et al. 2020). People typically inhale the microscopic spores when the environment in which Histoplasma fungi lives in, is disturbed (Centers for Disease Control and Prevention, 2021). When the microscopic spores in the mold form enter a human body, the fungus performs a dimorphic switch, turning into yeast (Gilmore et al. 2015; Beyhan & Sil 2019). In its yeast form Histoplasma capsulatum causes a disease called Histoplasmosis. Each year, up to 250,000 people in the U.S. are found to have Histoplasmosis (Fayyaz et al. 2020). The disease does not spread because of person-to-person contact (Stöppler 2020; Beyhan & Sil 2019).

Symptoms of Histoplasmosis include fever, cough, fatigue, chills, headache, chest pain, and body aches (Kauffman 2007). The spectrum of the infection includes mild, acute, chronic, and life-threatening sepsis (Kauffman 2007, Wheat et al. 2000). Although the disease is not easy to diagnose, the disease mostly goes away in a few weeks (Kauffman 2007, Knox & Hage 2010). When it becomes acute or chronic, antifungal medication is required (Knox & Hage 2010). Chronic pulmonary histoplasmosis if left untreated can have a mortality rate of 50% and with treatment it can have a mortality rate of 28% (Stöppler 2020).

Doctors diagnose Histoplasmosis through CT scans and X-rays (Stöppler 2020; Wheat et al. 2007). There is no vaccine against the disease. There are no drugs that specifically target the disease (Stöppler 2020; Wheat et al. 2000). Doctors recommend common antifungal medications of Itraconazole and amphotericin B for Histoplasmosis; patients with severe Histoplasmosis undergo treatment for several months (Wheat et al. 2000). Besides Histoplasmosis, Ocular histoplasmosis is most common in people who are exposed to the fungus at a very young age (U.S. Department of Health and Human Services, 2020). It happens when the fungus spreads from the lungs to the eyes. Early diagnosis and treatment are essential in preventing vision loss.

Existing State of Research

The Sil lab in the University of California - San Francisco (UCSF) has used mRNA sequencing and ribosomal footprinting to calculate the Translational Efficiency (TE) for each gene in four different strands of Histoplasma: G217B, HcH88, HcG186AR, and HcH143 (Gilmore et al. 2015). TE is measured as the footprint counts over mRNA counts. Researchers use TE as a key tool in measuring changes in RNA levels between different cell states. It is used as an indicator of protein production. By observing the sequence of the genes, the researchers have hypothesized that the immense variance in translational efficiencies for long and short DNA sequences could be because tRNA reads the stop codon prematurely after reading the start codon, which may result in the entire DNA message not getting translated accurately (Gilmore et al. 2015; Beyhan & Sil 2019). Even though the researchers noted several biomarkers that "may" influence TE, they have not found the actual list of biomarkers or sequence elements that best correlate with TE (Gilmore et al. 2015; Arribere & Gilbert 2013). Also, researchers have not figured out how they can use biomarkers to predict the TE computationally.

Goals

The goals of this research are the following:

- Identify sequence elements that affect TE in the yeast form of Histoplasma.
- Create a computational model to predict the TE in the yeast form of Histoplasma.

Methodology

Data from HistoBase database in the Sil Lab, UCSF was used in this study. Python code was developed to extract data from the database for 4 strains of Histoplasma, each with over 6500 genes. The 4 strains were G217B, HcH88, HcG186AR, and HcH143. Then, data was filtered to contain similar genes across 4 strains which resulted in 4891 genes. Subsequently, different sequence elements in the 5' Untranslated Region (UTR) were examined and its correlation to TE was determined using normal distribution curves, scatter plots, Wilcoxon tests, and Receiver Operating Characteristics (ROC) curves.

The following sequence elements were investigated to determine their correlation to TE:

1. Length of the Upstream Open Reading Frames (uORFs): The uORFs were chosen because in eukaryotic mRNAs, the translation of the protein requires the translation of uORFs (Gilmore et al. 2015). These biological structures aid in repressing or non-repressing the gene, which can lead to lower or higher

TE values respectively (Gilmore et al. 2015; Arribere & Gilbert 2013). The maximum length of the uORF was chosen because we hypothesized that greater length could have a higher influence on the initiation of translation. To get the largest length of the uORF, code was developed to extract the start and end points of the different uORFs, and their difference was taken for the 4981 genes. To determine if there is a correlation between TE and length of the uORFs, a scatter plot and a box plot were plotted. A Wilcoxon test was used to verify the results. Code was developed to calculate the true positive rate (TPR) and the false positive rates (FPR) using the data from the scatter and box plots and Receiver Operating Characteristic (ROC) curve was generated.

FPR = (False Positives)/ (Total Negatives)
= (False Positives)/ (False Positives + True Negative)

TPR = (True Positives)/ (Total Positives)
= (True Positives)/ (True Positives + False Negatives)

The ROC plot was analyzed to determine if TE was affected by length of the uORFs.

2. Secondary RNA structure: The energy of constrained secondary RNA structure was studied to see if secondary structures such as hairpin or loops could interfere with translation. A ROC curve using the energy of the constrained RNA structure against TE was plotted to determine if there was any relation between them.
3. Length of the 5' UTR: 5' UTR is the region directly upstream of the start codon. Previous studies show that genes that had a low length of 5' UTR sometimes had a low maximum length of uORFs, and those genes are translationally repressed (Gilmore et al. 2015). The length of the 5' UTR as a predictor of TE was evaluated by plotting a ROC curve.
4. CG to ATG ratio in uORFs: In the nucleotides, CG has 3 hydrogen bonds while AT has 2 hydrogen bonds. Since the RNA unzips the DNA by breaking Hydrogen bonds, the effect of CG to ATG ratio on TE was evaluated using a ROC curve.

After examining the individual sequence elements and their relation to TE, the effect of the combined sequence elements was used to predict TE. Derived features consisting of the sum of the top five uORFs and the ratio of the sum of the top five uORFs to the length of 5' UTR were also used as part of the combined input. Computational models were built using Linear Regression, Lasso Regression, Decision Tree Regression, and Random Forest which are four different supervised machine learning (ML) algorithms to evaluate how well the combined features contributed to predicting TE. High variance in the input data was mitigated using standard scaler function of Sci-kit learn Python library. The standard scaler transforms the data to ensure that the standard deviation of the data is one and the mean of the data is zero. Input data was randomly split into training, test, and validation sets in 70:20:10 ratio. 10-fold cross validation was used while training the data to build the computational models. ROC plots, root-mean-squared-error (RMSE), and r-squared score were used to compare the effectiveness of the different ML models to predict TE. Figure 1 shows the workflow for building the computational models.

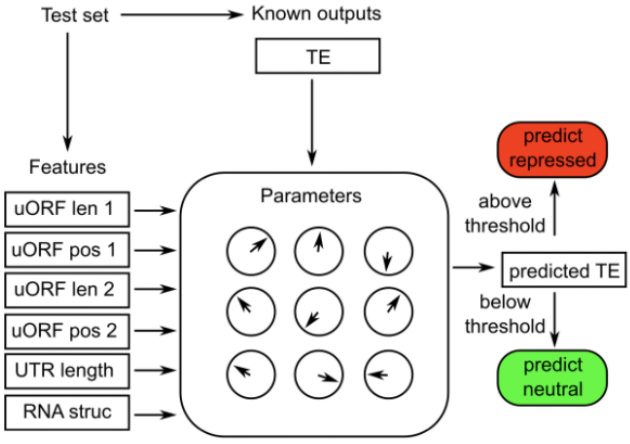

FIGURE 1. Workflow for building the computational models using Machine learning to

Results

The frequency distribution of TE values using a logarithmic scale is shown in Figure 2. The correlation between TE and length of the uORFs was visualized using a scatterplot as shown in Figure 3a. The scatter plot of maximum uORF lengths using ATG as the start codon vs. TE values is shown in Figure 3b. The Boxplot of the maximum length of a uORF across different genes is shown in Figure 4. The true positive (TPR) and the false positive rates (FPR) were calculated (Table 1) computationally using a threshold value of -2 and are plotted to generate a Receiver Operating Characteristic (ROC) curve in Figure 5a. The ROC plot of energy of the constrained RNA structure vs. TE is shown in Figure 5c. The ROC plot of the relationship between length of the 5' UTR and TE (in blue) using True Positive Rate (tpr) vs False Positive Rate (fdr) is shown in Figure 5d. The ROC plot of the relationship between CG/ATG ratio and TE (in green) using True Positive Rate (tpr) vs False Positive Rate (fdr) is shown in Figure 5e.

	Feature ≥ Threshold	Feature < Threshold
$\log_2(TE) < -2$	TP	FN
$-2 \leq \log_2(TE) < 2$	FP	TN

Table 1. The confusion matrix for the classification of TE prediction

To evaluate the quality of the results of the various ML algorithms, three standard metrics were used: ROC curves, root-mean-squared-error (RMSE), and r-squared score.

Linear regression gave a training RMSE of 1.65, a testing RMSE of 1.67, a testing r2 score of 0.23, and a training r2 score of 0.27. Figure 6a shows the coefficients of the features used in the linear regression algorithm.

Lasso regression gave a training RMSE of 1.63, a testing RMSE of 1.72, a testing r2 score of 0.16 and a training r2 score of 0.23. Figure 6b shows the coefficients used for Lasso regression algorithm. Figure 5g compares the ROC plots of the Lasso Regression with the maximum length of the uORF.

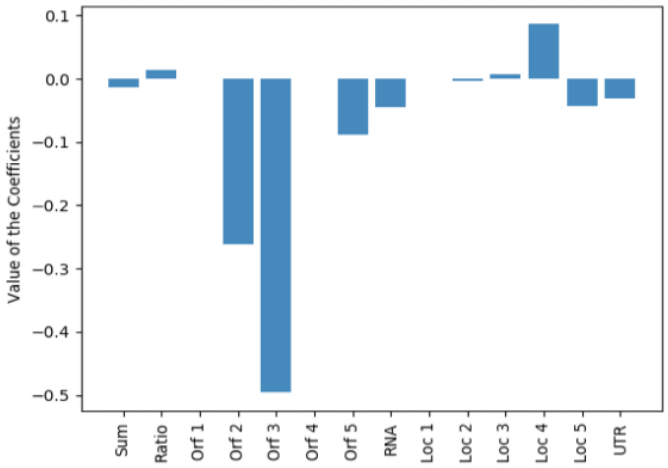

Figure 6a. Coefficient values of the features used in

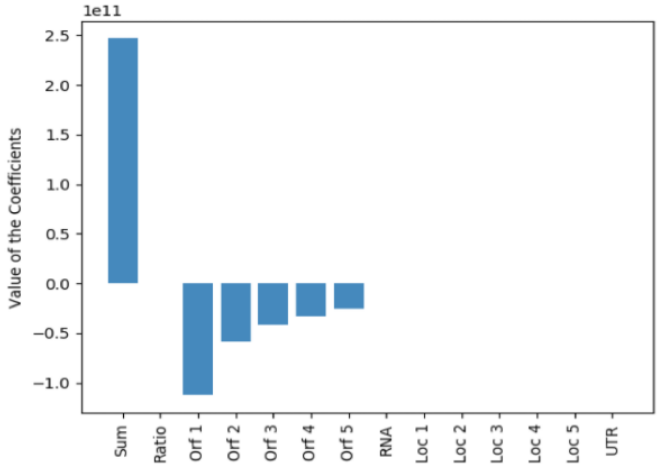

Figure 6b. Coefficient values of the features used Lasso Regression

The Decision Tree Regressor gave a training RMSE of 1.64, a testing RMSE of 1.79, a testing r2 score of 0.2, and a training r2 score of 0.23. Figure 5h compares the ROC plots of the Decision Tree to the maximum length of the uORF.

Random Forest gave a training RMSE of 1.37, a testing RMSE of 1.39, a testing r2 score of 0.47, and a training r2 score of 0.46. Figure 5i compares the ROC plots of Random Forest with the maximum length of the uORF. Figure 7 shows

the comparison of the various ROC plots with each other.

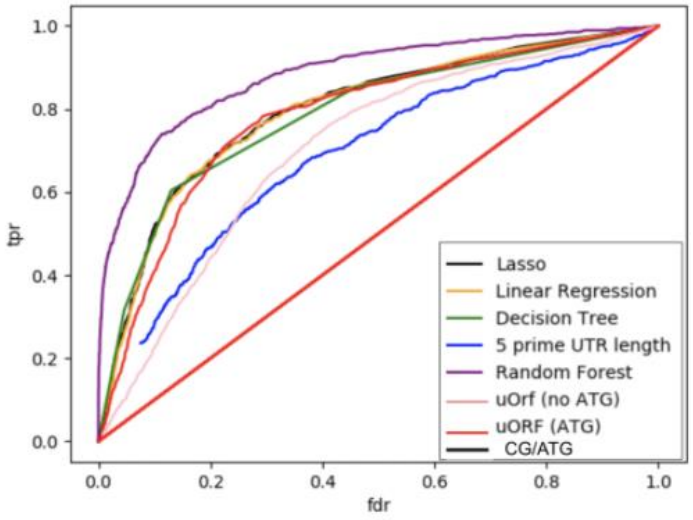

Figure 7. Relative comparison of the ROC plots for various learning models

Discussions

Figure 2 indicates that the frequency distribution of TE is not a normal distribution. Under ideal conditions, the number of footprint counts will be equal to mRNA counts but the graph indicates a spike in the values between -2 and 2.

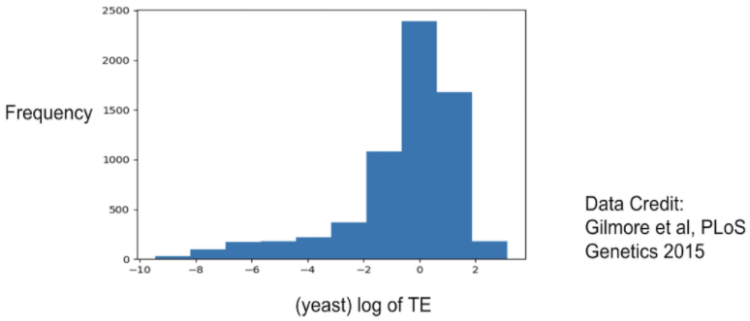

Figure 2. Frequency distribution of TE values

Upstream Open Reading Frames (uORFs) vs. TE

Figure 3a depicts a weak negative correlation between the TE values and maximum length of the uORFs, where each dot represents a gene common across all four strains of Histoplasma: HcG217B, HcH88, HcH143, HcG186Ar. Figure 4 conveys that the length of the largest uORF indeed significantly distinguishes translationally repressed genes from neutral genes. This was verified through the Wilcoxon test, in which the P-Value was 2.2e - 16. Based on Figure 3a and Figure 4, a threshold value of -2 was chosen to generate a Receiver Operating Characteristic (ROC) curve in Figure 5a.

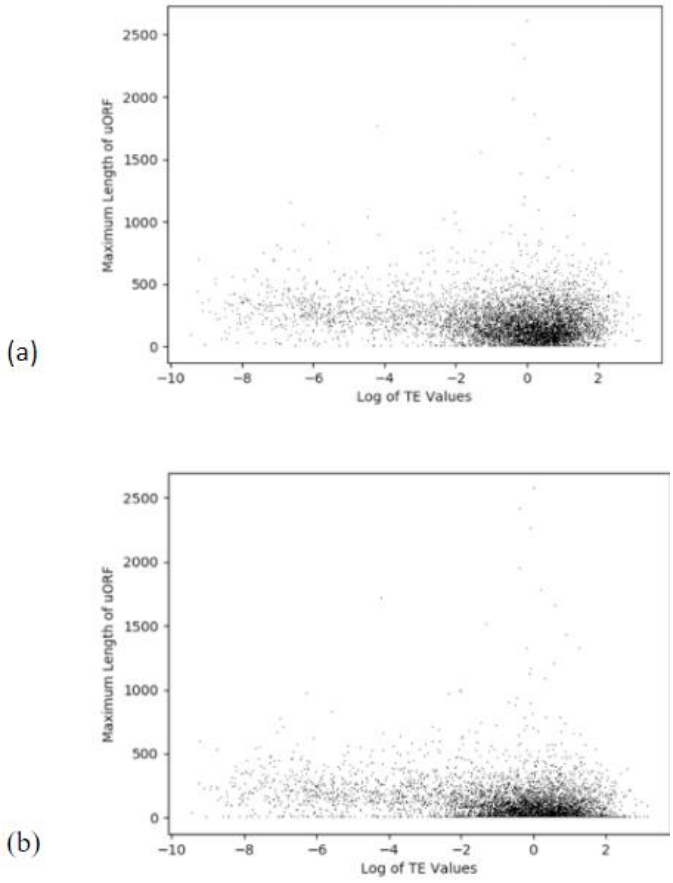

Figure 3. Scatter plot of maximum uORF lengths vs TE values for a) with any start codon b) using ATG as the start codon

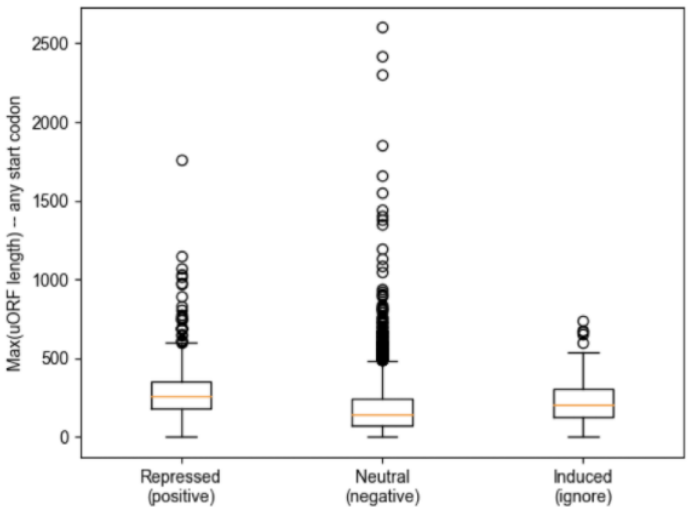

Figure 4. Boxplot of the maximum length of a uORF across different genes

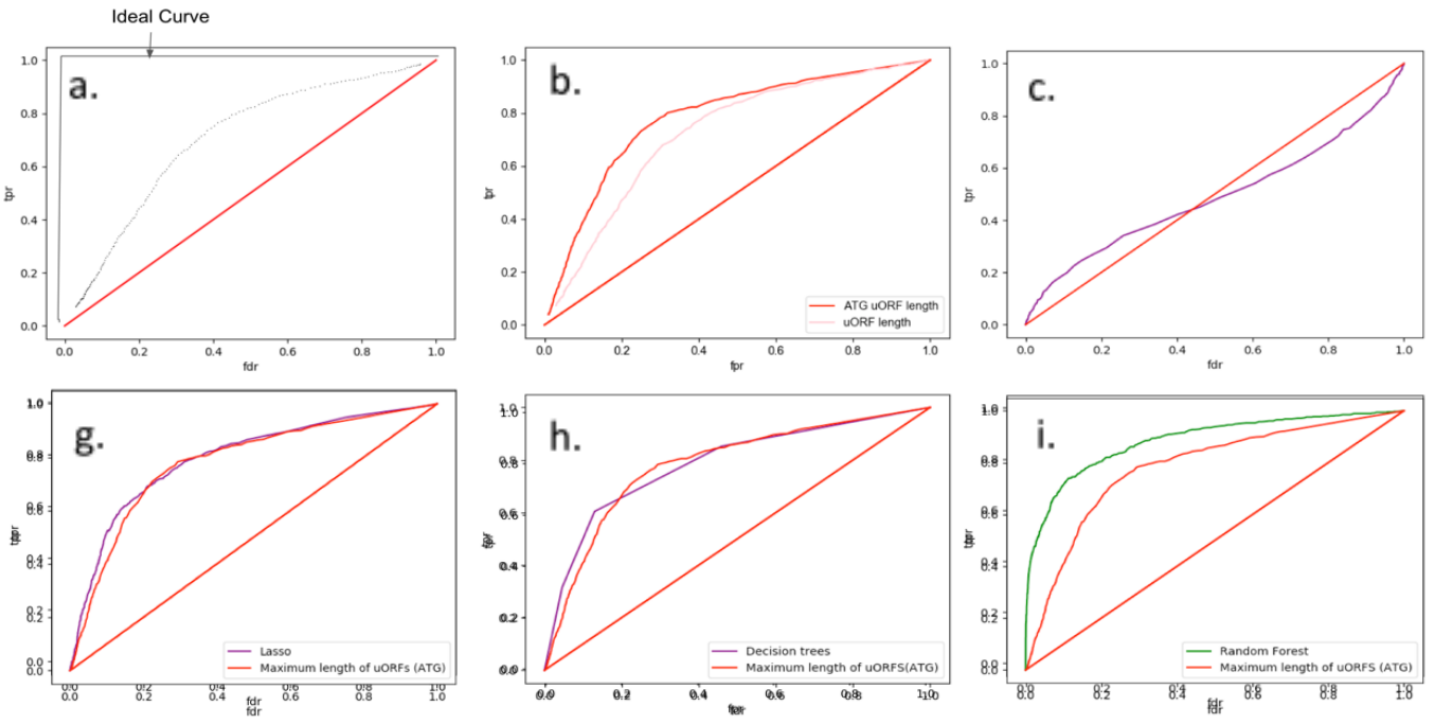

Figure 5. ROC plot showing the relationships between a feature/predictor and TE | a) maximum uORF length for any start codon (in black dots) | b) maximum uORF length with ATG start Codon (in red) | c) energy of constrained secondary RNA structure (in purple) | d) length of the 5' UTR and TE (in blue) | e) CG/ATG ratio (in green) | f) Linear regression (in green) | g) Lasso regression (in purple) | h) Decision Tree Regressor (in purple) | i) Random Forest Regressor (in green)

Analysis of the ROC plot of the maximum length of uORF vs. TE has area under the curve (AUC) of 0.68 which predicts TE better than mean value. It shows that the maximum length of uORF works as a good predictor of TE (i.e., the genes that were repressed have a correlation with maximum length of uORFs). However, in this case any start codon was used when calculating the maximum length of the uORF. ATG (i.e. AUG) is the start codon of mRNA as it is first to undergo translation after transcription. When the analysis was repeated using the ATG start codon (vs. the use of any start codon in the previous analysis), it was observed that there still existed a weak negative correlation between the open reading frames and the translational efficiency (Figure 3b). However, the ROC for this scenario with only ATG as start codon, revealed a bigger area under the curve with AUC = 0.74 (Figure 5b) compared to the previous ROC curve. Therefore, the maximum length of uORF with start codon ATG correlates better to TE.

Secondary RNA structure vs. TE

The ROC plot (Figure 5c) of energy of the constrained RNA structure vs. TE showed that RNA structure does not correlate well with TE. In fact, the second half of the ROC plot conveys that the probability of predicting TE is less than the probability of randomly flipping a coin.

Length of 5' UTR vs. TE

In the ROC plot (Figure 5d) of the length of 5' UTR vs. TE showed that the curve of the uORFs with start codons of "ATG" (in red) performs better than the length of the 5' UTR (in blue). Nevertheless, the correlation between TE and length of 5' UTR with AUC = 0.63 was better than the average value.

CG to ATG ratio in uORFs

The ROC plot (Figure 5e) with AUC = 0.79 shows that there is a very good correlation between the CG/ATG ratio (in green) and TE and it performed better than uORFs with start codons of "ATG" (in red).

Linear Regression

From Figure 6a, we can infer that low R-squared values are due to high variance in the data. The coefficients also indicate that the sum of the top five uORFs as well as the maximum length of the uORF per gene has a high correlation with TE.
Figure 5f compares the ROC plots of the Linear Regression with the maximum length of the uORF. The ROC plot shows that Linear regression using all the features does slightly better with AUC = 0.75 than just using the maximum length of the uORF (AUC = 0.74).

Lasso Regression

Since there is some difference between both the RMSE and the r-squared scores, Lasso model slightly overfits the data. Contrary to the Linear regression model, the sum of the top five uORFs and the maximum length of the uORF did not have a high coefficient value. Instead, the second and third largest uORF lengths had the highest coefficient values. The ROC plot in Figure 5g shows that Lasso regression (AUC = 0.74) performs the same as the maximum length of the uORF and does not add any better performance compared to Linear regression.

Decision Tree

The algorithm slightly overfits because of the difference between the RMSE scores for training and testing. The ROC plot in Figure 5h shows that the Decision Tree (AUC = 0.69) performs worse than the maximum length of the uORF predictor.

Random Forest

Unlike the previous algorithms, Random Forest performed substantially better with AUC = 0.85 (Figure 5i) compared to the maximum length of the uORF predictor.

Comparing all results

The sequence elements CG to ATG ratio and maximum length of uORF with ATG as the start codon, each by themselves alone, gave the best performance with AUC at 0.79 and 0.74 respectively and therefore demonstrated the best correlation with TE for single features. The uORF with ATG as the start codon acts as a better predictor than uORF with any start codon because the former (relatively speaking) potentially increases the likelihood of translation starting before it reaches the main functional ORF. The length of 5' UTR to predict TE (AUC = 0.63) was not a good predictor of TE compared to the above two sequence elements. The biological reason why the length of 5' UTR predictor is possibly worse than the largest uORF predictor could be that the length of a UTR and the maximum length of uORF are not well correlated with each other, i.e. you can have a long 5' UTR region with many small length uORFs in it. The Secondary RNA structure was the poorest predictor of TE possibly because the positions of the hairpins are probably not ideal to influence the initiation or repression of translation.

The different learning models can be assigned to the categories of 1) Single feature (uORF-no ATG, 5' UTR length, uORF - ATG) 2) Multiple features (Decision Tree, Linear, Lasso) and 3) Ensemble of multiple features (Random Forest).

The RMSE test and train values for the different computational models built using the combined sequence elements to predict TE were comparable and had little to no overfitting. The low r-squared values in all models indicated that the data has high variance. Lasso regression model (AUC = 0.74) and Linear regression model (AUC = 0.75) performed comparably to the maximum length of uORF with ATG as the start codon (AUC = 0.74) and CG to ATG ratio (AUC = 0.79). Decision trees model performed slightly worse with AUC of 0.69. Random forest model performed best with AUC at 0.85 and had lowest RMSE of 1.37. The ensemble method of Random

Forest works better than the individual Decision trees model because the individual model tends to make different types of errors on the subset of features chosen, and many of those errors cancel each other out in the ensemble method (Géron, 2020). Overall, the combination of multiple features, except when using Decision Trees, yielded slightly better results compared to using single features while predicting TE. This is to be expected as a model with multiple features has more information in building the decision boundary of a classifier.

The features and methods used in the prediction of TE in Histoplasma capsulatum can be applied to other single-cell eukaryotic organisms to improve our insight into how the sequence elements in the 5' UTR affects translation in other eukaryotic organisms. This would also make the findings from this research useful in the research of other diseases, not just Histoplasmosis.

Conclusions

This research was the first to discover two sequence elements, CG to ATG ratio and maximum length of uORF with ATG as the start codon, in 5' UTR of mRNA that affect the TE in Histoplasma capsulatum. This research was also the first to develop four computational models for predicting TE of Histoplasma using the combined sequence elements. The recommendation is to use the computational model developed using Random Forest for predicting TE. This research improves our understanding of the sequence elements in the 5' UTR affecting the TE of Histoplasma capsulatum. In the future, similar features and methods can be used to predict the TE in other single-cell eukaryotic organisms which can improve our understanding of the translation process in other organisms. Also, in the future, the computational methods used in this research can be extended to use deep learning methods and clustering methods, and their results can be compared with existing models to see if they can improve the predictions.

Acknowledgments

I would like to sincerely thank Mark Voorhies, Ph.D., Department of Microbiology and Immunology, University of San Francisco School of Medicine, for his invaluable mentorship and Dr. Anita Sil, MD, Ph.D., Professor of Microbiology and Immunology, University of San Francisco School of Medicine, for giving me the opportunity to perform this research at the SIL Lab and for permitting me to use the data from HistoBase for this project.

References

Gilmore, S. A., Voorhies, M., Gebhart, D., & Sil, A. (2015). Genome-wide reprogramming of transcript architecture by temperature specifies the developmental states of the human pathogen histoplasma. PLOS Genetics, 11(7). https://doi.org/10.1371/journal.pgen.1005395

Inglis, D. O., Voorhies, M., Hocking Murray, D. R., & Sil, A. (2013). Comparative transcriptomics of infectious spores from the fungal pathogen histoplasma capsulatum reveals a core set of transcripts that specify infectious and pathogenic states. Eukaryotic Cell, 12(6), 828–852. https://doi.org/10.1128/ec.00069-13

Bahr, N. C., Antinori, S., Wheat, L. J., & Sarosi, G. A. (2015). Histoplasmosis infections worldwide: Thinking outside of the Ohio River Valley. Current Tropical Medicine Reports, 2(2), 70–80. https://doi.org/10.1007/s40475-015-0044-0

Centers for Disease Control and Prevention. (2021, January 14). About histoplasmosis. Centers for Disease Control and Prevention. https://www.cdc.gov/fungal/diseases/histoplasmosis/definition.html

Beyhan, S., & Sil, A. (2019). Sensing the heat and the host: Virulence determinants of Histoplasma capsulatum. Virulence, 10(1), 793–800. https://doi.org/10.1080/21505594.2019.1663596

Fayyaz, J., Vydyula, R., Walczyszyn, M. P., & Lessnau, K.-D. (2020, December 6). What is the incidence of histoplasmosis in the US? Medscape. Retrieved May 6, 2020, from https://www.medscape.com/answers/299054-108391/what-is-the-incidence-of-histoplasmosis-in-the-us

Stöppler, M. C. (2020, December 30). Histoplasmosis (cave disease): Treatment, symptoms & diagnosis. MedicineNet. Retrieved from July 2, 2020, from https://www.medicinenet.com/histoplasmosis_facts/article.htm

Kauffman, C. A. (2007). Histoplasmosis: A clinical and Laboratory update. Clinical Microbiology Reviews, 20(1), 115–132. https://doi.org/10.1128/cmr.00027-06

Wheat, J., Sarosi, G., McKinsey, D., Hamill, R., Bradsher, R., Johnson, P., Loyd, J., & Kauffman, C. (2000). Practice guidelines for the management of patients with histoplasmosis. Clinical Infectious Diseases, 30(4), 688–695. https://doi.org/10.1086/313752

Knox, K. S., & Hage, C. A. (2010). Histoplasmosis. Proceedings of the American Thoracic Society, 7(3), 169–172. https://doi.org/10.1513/pats.200907-069al

Wheat, L. J., Freifeld, A. G., Kleiman, M. B., Baddley, J. W., McKinsey, D. S., Loyd, J. E., & Kauffman, C. A. (2007). Clinical practice guidelines for the management of patients with histoplasmosis: 2007 update by the Infectious Diseases Society of America. Clinical Infectious Diseases, 45(7), 807–825. https://doi.org/10.1086/521259

U.S. Department of Health and Human Services. (2020, November 20). Ocular histoplasmosis syndrome (OHS). National Eye Institute. Retrieved July 8, 2020, from https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/ocular-histoplasmosis-syndrome-ohs

Arribere, J. A., & Gilbert, W. V. (2013). Roles for transcript leaders in translation and mrna decay revealed by transcript leader sequencing. Genome Research, 23(6), 977–987. https://doi.org/10.1101/gr.150342.112

Géron, A. (2020). Ensemble methods. In Hands-on machine learning with scikit-learn, Keras, and tensorflow: Concepts, tools, and techniques to build Intelligent Systems (2nd ed., pp. 74–75). O'Reilly.