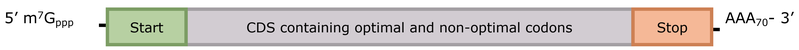
Understanding the Molecular Origins of Transcriptome Degradation Using Machine Learning

Judith Rodriguez, M.S. [1,2], Justin Petucci Ph.D. [4], Vasant Honavar, Ph.D [3,4] ,Edward O'Brien, Ph.D. [1,2,4]

1) Huck Institutes of Life Sciences, Pennsylvania State University, University Park, 16802 2) Department of Chemistry, Pennsylvania State University, University Park, 16802 3) College of Information Sciences and Technology, Pennsylvania State University, University Park, 16802 4) Institute for Computational and Data Sciences, Pennsylvania State University, University Park, 16802

Translational optimality of transcripts can influence ribosome movement and determine mRNA half-life as a result. (*Presnyak et al 2015; Sharma and O'Brien 2019*)



The enrichment and placement of non-optimal codons promotes ribosome queues causing ribosomal congestion on the transcript and is recognized by translation-dependent degradation. (*Presnyak et al 2015; Sharma and O'Brien 2019*)

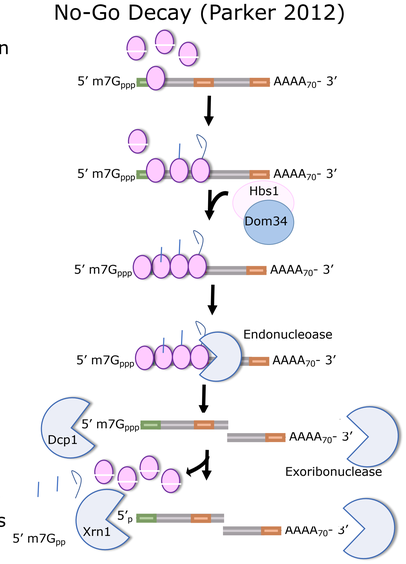
Over 6000 transcripts with half-lives ranging between 0.7 and 38 min exist in S. cerevisiae. (*Chan et al 2018*)

Previous prediction models have resulted in moderate variability (*Neymotin et al 2013; Cheng et al 2018*)

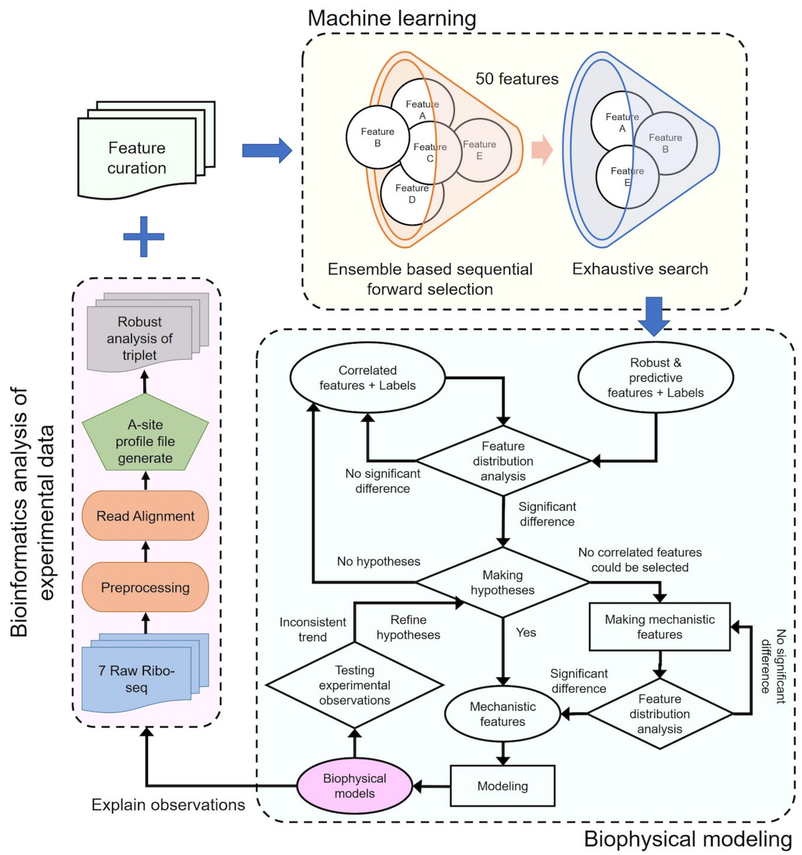
Our features are based on kinetic information from codon translation rates.

Obtaining these preliminary results is the first step in simulating mRNA degradation and experimentally testing our Machine Learning approach to offer novel insights into the molecular factors governing transcriptome half-life.

No-Go Decay (Parker 2012)



Big data and machine learning are used to identify the most robust and predictive features out of 383 features for a biophysical model that describes mRNA degradation



Robust and Predictive Features identified from Machine Learning

Figure 1: Negative MSE score improves with increased number of feature selections

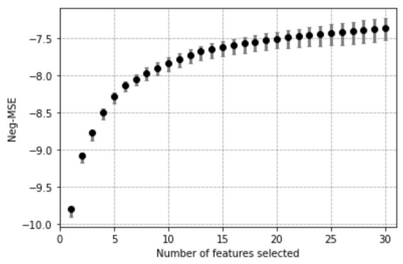


Figure 2: Top 30 features were identified by Selection Metric calculated for 2000 iterations

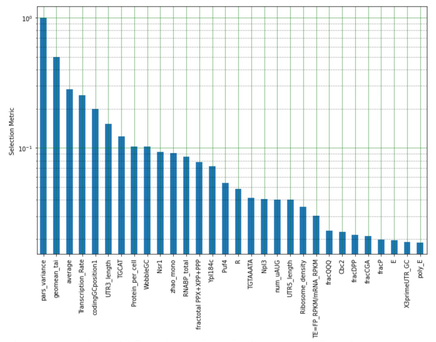


Figure 3: Robust and predictive features identified calculating a feature occurrence ratio across models of 4 to 6 features

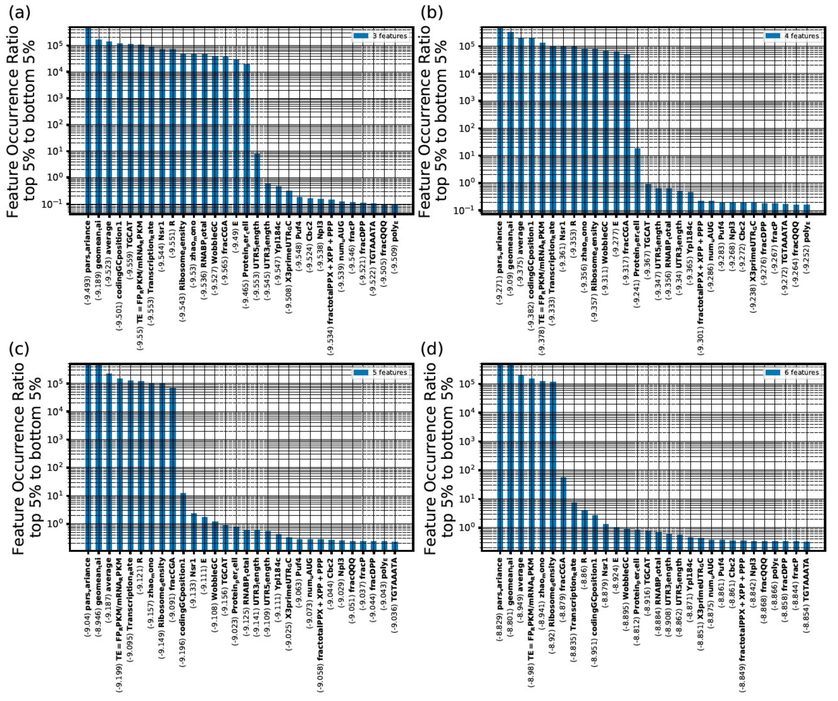


Figure 4: 10 Features are shared across 5/6 Robust Predictive Feature Models

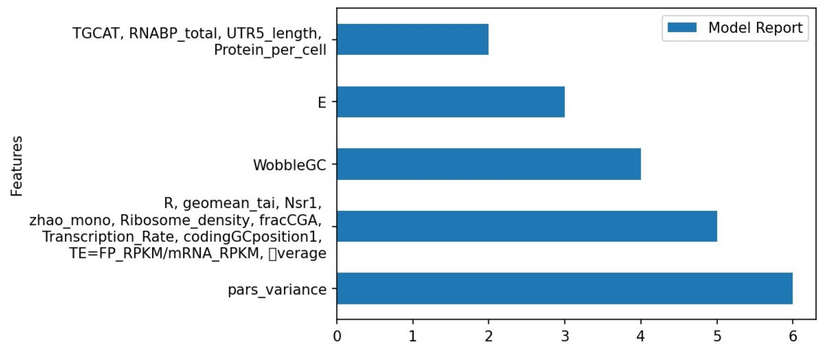
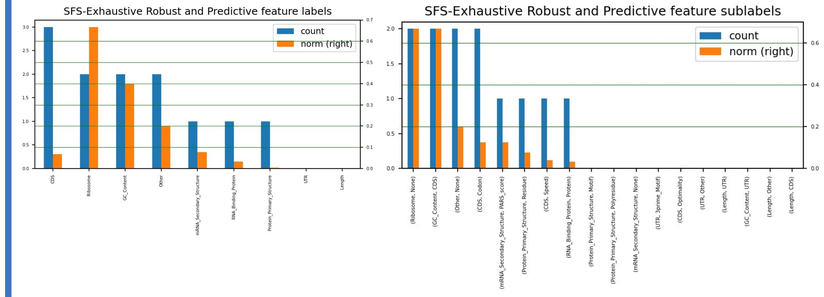


Figure 5: Feature label analysis shows most of the robust and predictive features are derived from coding sequence characteristics.



Example of feature distribution and correlation analysis to generation a hypothesis using the Feature R (the total amount of arginine in a transcript)

Figure 6: Use 25% and 75% quantiles of half-life distribution to analyze half-life extremes

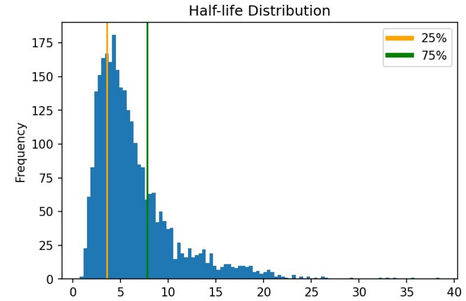
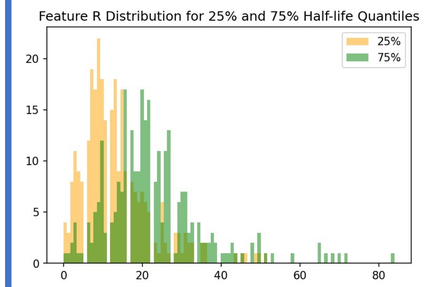
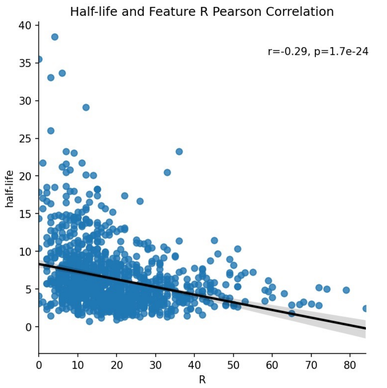


Figure 7: Half-lives are shorter with increased amount of arginine in the protein sequence





Arginine is a positive charged residue that interacts with the negatively charged ribosomal exit tunnel and in result can slow down translation which can trigger ribosomal collisions.

If the ribosomal collision is not rescued, the No Go Decay pathway can be triggered, and degradation of the transcript begins.

Hypothesis: Transcripts enriched with arginine in their sequence slow down translation and have increase decay rates.

Future Work

Generate hypothesis for remaining robust and predictive features.

Create biophysical model from of these generated hypothesis'

Experimentally test hypothesis of robust and predictive features and biophysical model

NIH National Institutes of Health
Office of Extramural Research