Model of Translation in *Plasmodium falciparum*

Anmol Kagrecha, EE, IITB
Chhaminder Kaur, BSBE, IITB
Swati Patankar, BSBE, IITB
Narayan Rangaraj, IEOR, IITB

1 Introduction

The genome of *P. falciparum* is highly AT-rich with AT content as high as 90% in the intergenic regions leading to many more upstream start codons and upstream open reading frames (uORFs) on the messenger RNAs compared to other eukaryotes. An open reading frame (ORF) is a continuous stretch of codons that begins with a start codon and ends at a stop codon.

Under the assumption that the ribosome scans the mRNA linearly, the ribosome will encounter many uORFs before reaching the coding region (CDS). Using data from a few in vitro experiments, we constructed a model which could enable us to understand the translation over all the genes in *P. falciparum*.

However, our model wasn't able to generalize beyond the experimental data that we had. We have realized that there are many more factors involved in the translation machinery of *P. falciparum* than we're able to account for. The details of our model and its performance are given in the subsequent sections.

2 Preliminaries

Some of the known factors affecting the translation of any ORF are:

- mRNA cap: Ribosome with the help of initiation factors recognizes the mRNA cap and scans the mRNA searching for start codons to initiate translation.
- Kozak sequence: It is the sequence flanking the start codon (AUG). It is usually taken as a sequence of 5 bases upstream of AUG and 1 base downstream of AUG.

  The upstream locations are denoted as -5 to -1 positions in the Kozak sequence. The A in AUG is denoted as the +1 position and the downstream location is denoted by +4 position. Note that there is no location which is denoted by 0.
- Codon usage: Sequences of the codons making up the ORF, determined by the genetic code of 61 codons giving information for 20 amino acids. From the numbers, it is clear that multiple codons can code for the same amino acid. Codon usage refers to the relative frequencies in the coding sequences of the codons coding for the same amino acid.

3 The Translation Model

We're interested in finding the probability of translation of the coding sequence in a gene. To find this probability, we divide the translation process into the following sub-processes:

- Initiation: Initiation is the process in which the ribosome starts translation of an ORF. The probability of initiation is denoted by p_{init}.
- Elongation: Elongation is the process in which the ribosome acquires amino acids for the codons in mRNA. If there are codons for which there are less amino acids, the ribosome may stall and not form a protein. The probability of completing elongation is denoted by p_{elon}.
- Dissociation: After completing the elongation, the ribosome may dissociate completely or keep scanning for new ORFs. The probability of not dissociating is denoted by p_{nd}.

- Reinitiation: If the ribosome doesn't dissociate completely and keeps scanning, it has to acquire met-tRNA to be able to initiate translation again. We denote the probability of acquiring met-tRNA as p_{reinit}.

3.1 Probability of Initiation

Analysis on data from our lab [KSP15] was performed to find the most important position in Kozak Sequence that affects translation. Clustering sequences according to their +4 positions gave the least standard deviation compared to clustering with other positions. Confidence intervals with confidence of 90% for initiation probabilities are:

- A at +4 position: [0.127, 0.151] and estimate = 0.139
- T at +4 position [0.504, 0.619] and estimate = 0.562
- G at +4 position [0.647, 0.812] and estimate = 0.729
- C at +4 position [0.291, 0.437] and estimate = 0.364

Wilson score interval [Wal13] was used to estimate the probabilities and the confidence intervals.

3.2 Probability of Elongation

We have used codon adaptation index (CAI) [SL87] to account for codon usage bias. The CAI is a geometric mean of the weight associated to each codon over a length of the ORF sequence. CAI is defined as:

$$\text{CAI} = (\prod_{1}^{L} w_i)^{\frac{1}{L}},$$
$$w_i = \frac{f_i}{\max f_j}$$

where, L is the number of codons in the ORF, w_i is the weight associated with each codon, f_i is the frequency of occurrence of each codon and most importantly, $[i, j]$ are synonymous codons for an amino acid, i.e., they code for the same amino acid.

Codon usage table for *P. falciparum* was retrieved from NCBI-GenBank. In particular we used this website. The weights for the CAI were calculated according to the codon usage table.

3.3 Probability of Not Dissociating

It was observed experimentally that as the length of an ORF increases, the probability of not dissociating becomes very small. Ceteris paribus, a larger uORF would have a larger repressing effect on the translation of the coding sequence. An exponentially decaying function was chosen to fit the limited data we have. The value of the function w.r.t. length of ORF are plotted in Figure 1.

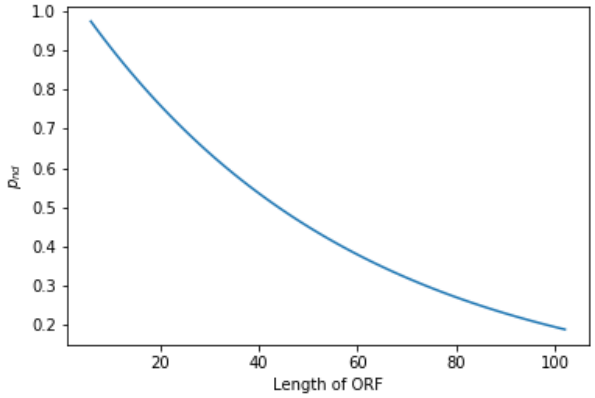

Figure 1: p_{nd} vs. Length of ORF

3.4 Probability of Reinitiation

The base pairs between the stop codon of one ORF and the start codon of another ORF is called the intercistrionic length between the two ORFs. Very short intercistrionic lengths do not allow sufficient time to acquire met-tRNA, which make it difficult for the ribosome to initiate again. A shifted logistic function was used to fit the limited data we have. The value of the function w.r.t intercistrionic length are plotted in Figure 2.

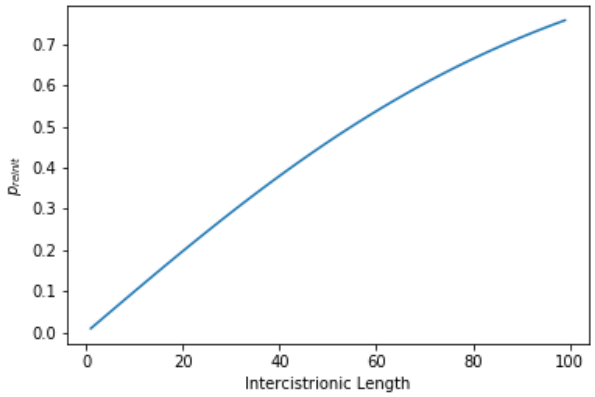

Figure 2: p_{reinit} vs. Intercistrionic Length

3.5 Probability of Translation

We take a Markovian approach to find the probability of translation, i.e., the state in which a ribosome will be, depends only on the previous state. First, consider the simple example when there is a single uORF and a CDS. The possible paths that are followed by a ribosome with a single uORF are shown in Figure 3. The probability of translation can be written as:

$$p_{tran} = p_{\text{init2}} p_{\text{elon2}} (1 - p_{\text{init1}} + p_{\text{init1}} p_{\text{elon1}} p_{\text{nd1}} p_{\text{reinit12}})$$

where p_{init1}, p_{elon1}, and p_{nd1} are initiation, elongation and not dissociating probabilities corresponding to uORF, p_{init1} and p_{elon1} are initiation and elongation probabilities corresponding to CDS and p_{reinit12} is the reinitiation probability corresponding to uORF and CDS.

When there are multiple uORFs, the probability calculations can be done using Figure 4 and Figure 5. Figure 4 is a compact representation of an extended chain containing several uORFs and helps in calculating the probability of reaching the CDS. Figure 5 is a simple calculation of finding probability of translation given that the ribosome has reached the CDS.

4 Simulations

4.1 Correlation between Translational Efficiency and Probability of Translation

Translational efficiency is defined as the ratio of the protein produced and the mRNA in the system. The translational efficiency can be greater than 1 because a single mRNA could be translated multiple times before it decays. Translational efficiencies for 92 genes were calculated from the data in the lab.

The coefficient of determination R^2 between the predicted probability of translation and translational efficiency came out to be negative. Rather than comparing the values directly, we also compared the ranks of genes when sorted according to probability of translation and translational efficiency. The Spearman's rank correlation [FHP08] coefficient also came out to be negative.

4.2 Predicting Repressing Nature of uORFs

We were interested to find the effect of individual uORFs on the probability of translation. The computer program was suitably modified to neglect the effect of an uORF and find the probability of translation. An increase in probability of translation would indicate a repressing role of the uORF and vice-versa.

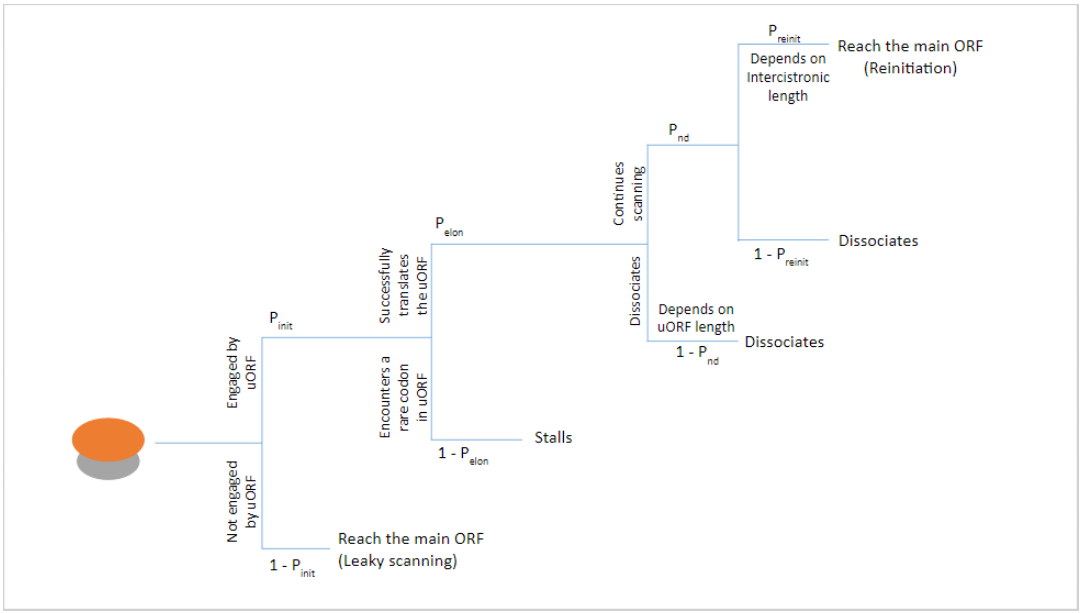

Figure 3: State model for a single uORF

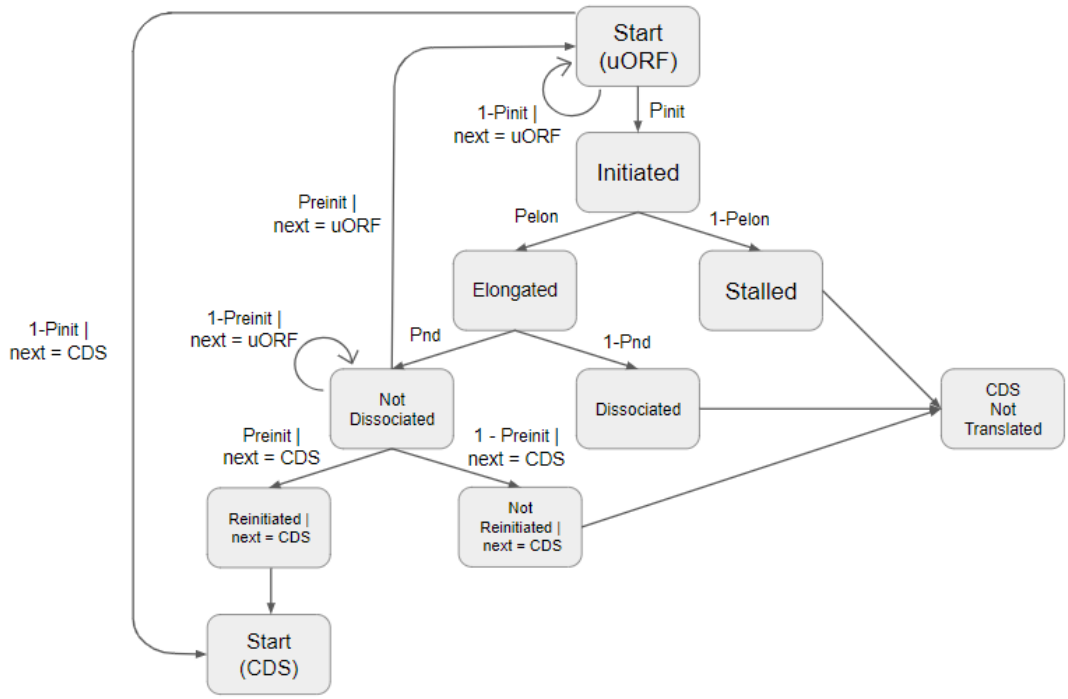

Figure 4: State model for multiple uORFs, part 1

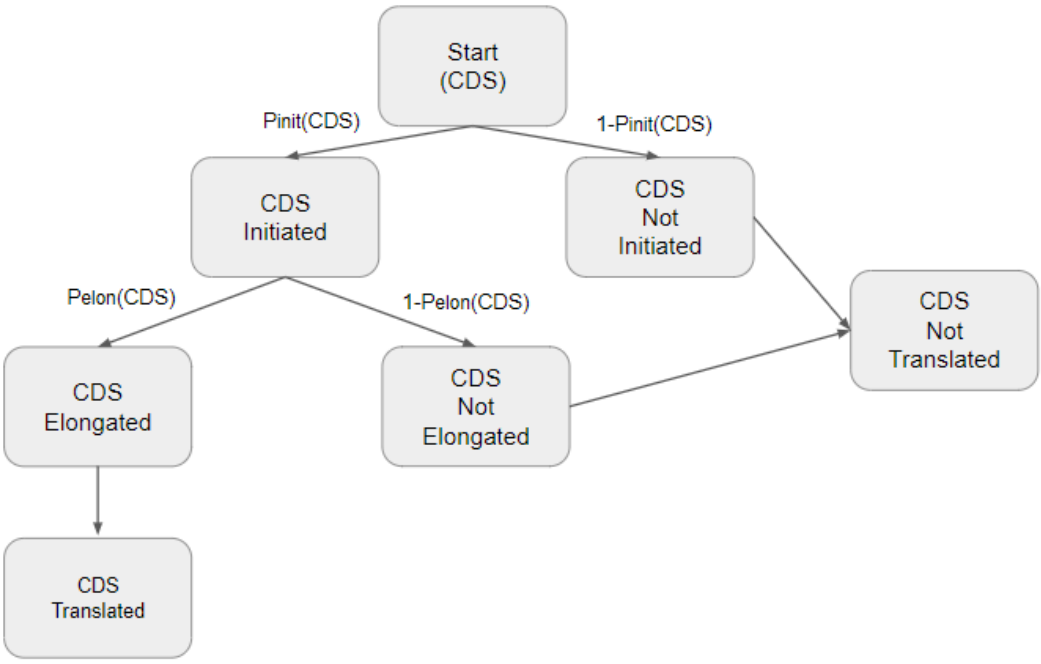

Figure 5: State model for multiple uORFs, part 2

We tested our procedure on the gene for protein VAR2CSA. VAR2CSA has a strongly repressing uORF at 270 base pairs before the CDS [Amu+09]. Our method predicted a non-repressing role of the uORF. The repressing uORF is a large uORF and contains 13 smaller uORFs. Ignoring this uORF for probability of translation calculations introduces the effects of the numerous smaller uORFs and makes the overall probability of translation smaller.

5 Conclusion

The simulations suggest that there are many other unknown factors that are required to be incorporated into the model for better predictive power. Instead of model based approaches, machine learning based approaches could be used if there is a large amount of data available.

References

[SL87] Paul M Sharp and Wen-Hsiung Li. "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications". In: *Nucleic acids research* 15.3 (1987), pp. 1281–1295.

[FHP08] E. C. Fieller, H. Hartley, and E. S. Pearson. *TESTS FOR RANK CORRELATION COEFFICIENTS. I.* 2008.

[Amu+09] Borko Amulic et al. "An upstream open reading frame controls translation of var2csa, a gene implicated in placental malaria". In: *PLoS pathogens* 5.1 (2009), e1000256.

[Wal13] Sean Wallis. "Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods". In: *Journal of Quantitative Linguistics* 20.3 (2013), pp. 178–208.

[KSP15] Mayank Kumar, Vivek Srinivas, and Swati Patankar. "Upstream AUGs and upstream ORFs can regulate the downstream ORF in Plasmodium falciparum". In: *Malaria journal* 14.1 (2015), p. 512.