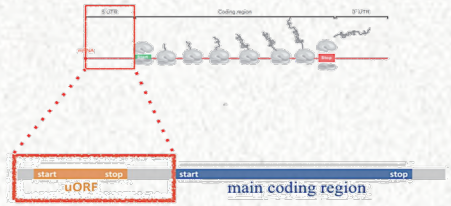
Hidden Markov Models for identification of translational regulatory elements

Jiayuan Zhu with supervisor Dr. Heejung Shim

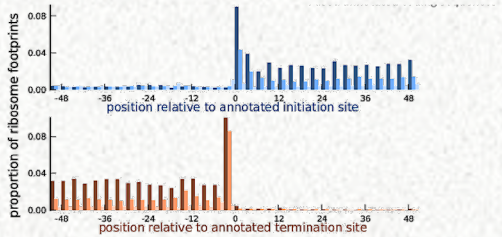
Introduction

When, where, and how much of a protein is expressed is important to explain the underlying biological processes. A major determinant of protein expression level is the rate of protein translation. One interesting regulatory element of the translation is the upstream Open Reading Frames (uORFs). These are relatively shorter sequences of codons, located upstream of the protein coding regions. Translation at uORFs has been known to affect translation of coding regions during important biological processes [1]. The methods introduced by Raj et al. [2], although originally designed for another purpose, have shown a promising start to the identification of translated uORFs. This project will build up on their work and aim to develop and implement statistical methods that enable more comprehensive identification of translated uORFs in tissues/organisms of interest.



Data - Ribosome Footprint Profiling

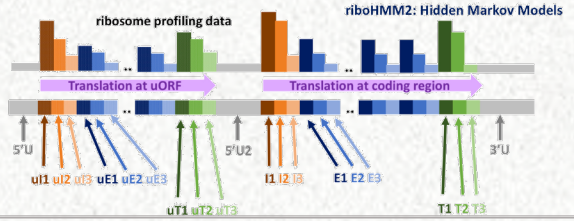
Ribosome profiling, a technique for directly quantifying levels of translation, has enabled genome-wide identification of translated regions [2]. The model is then built on the characteristics of the data. This figure illustrates that translation region has two typical features. The translated region tends to have significant higher abundance of ribosome footprints. Besides, the 3-base periodicity pattern exists in the translation region (known as codon).



Hidden Markov Model - Hidden States

According to the data characteristics, hidden Markov model is designed for this project. There are 21 hidden states:

- 5'U: 5' untranslated state
- uI1, uI2, uI3 : translation initiation at uORF
- uE1, uE2, uE3: translation elongation at uORF
- uT1, uT2, uT3: translation termination at uORF
- 5'U2: untranslated bases downstream of uORF
- I1, I2, I3: translation initiation at coding region
- E1, E2, E3: translation elongation at coding region
- T1, T2, T3: translation termination at coding region
- 3'U: 3' untranslated state

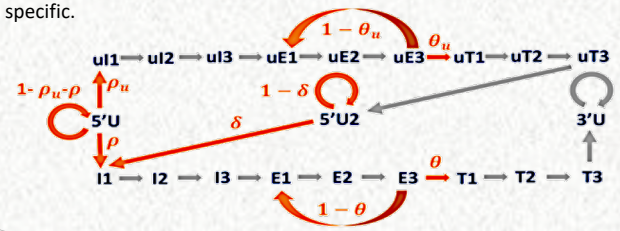


Hidden Markov Model - Transition Probability

It is assumed that:

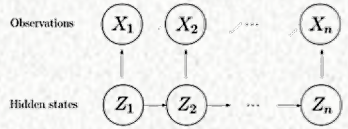
- The RNA sequences always start from 5'U
- Translation initiation can occur at non-canonical start codon
- Translation termination occurs at the first canonical stop codon (i.e. UAA, UGA, UAG)

Although there are 21 possible hidden states, most movements between states are deterministic. The grey arrows in the figure indicate that the transition probability is 1. θ_μ and θ would become 1 if the stop codon occurs. ρ_μ is related to the start for translation of uORF. ρ, δ represent the beginning for main coding region from 5'U and 5'U2, respectively. It is also clear that each start codon (e.g. AUG) would have different transition probabilities. So ρ_μ, ρ and δ are codon specific.



Hidden Markov Model - Emission Probability

Consider an RNA transcript that has length n. $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ represents the ribosome footprint counts and $\mathbf{Z} = (z_1, z_2, \ldots, z_n)$ denotes the hidden states for the n^{th} position in the transcript.



The emission distribution is modeled as follows:

$$x_n | \ z_n = z \ \sim \text{Poisson}(\lambda_{z,n})$$
$$\lambda_{z,n} \sim \text{Gamma}(\alpha_z, \beta_z)$$

The above formulation of the emission distribution of $x_{z,n}$ is in a hierarchical form. By integrating over the random Poisson rate, the equivalent single emission distribution of $x_{z,n}$ can be achieved in Negative Binomial form.

Parameter Estimation & Identification of Translational Regulatory Elements

- Compute the maximum likelihood estimates of the model parameters using Expectation-Maximization algorithm
- Infer the maximum of a posteriori (MAP) hidden state sequences using the Viterbi algorithm, and then identify translated uORFs using them

Ongoing Work

I'm currently implementing the proposed model with Python to identify translated uORFs in tissues/organisms of interest.

Reference

[1] Barbosa, C et al. 2013. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. PLoS Genetics

[2] Raj, A et al. 2016. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. eLife