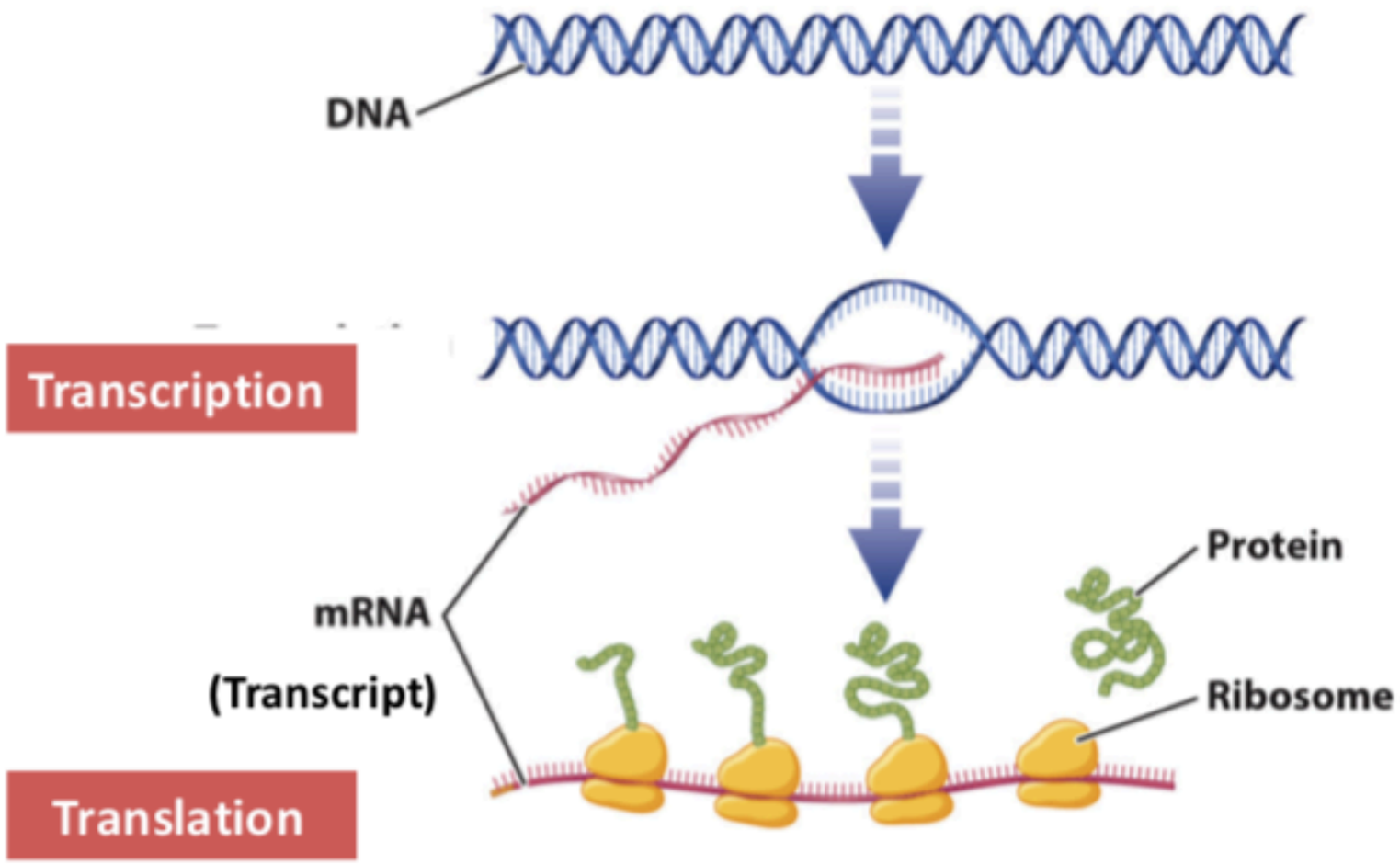
Hidden Markov Models for Identification of Translational Regulation Elements

Jiayuan Zhu
Supervisor: Heejung Shim

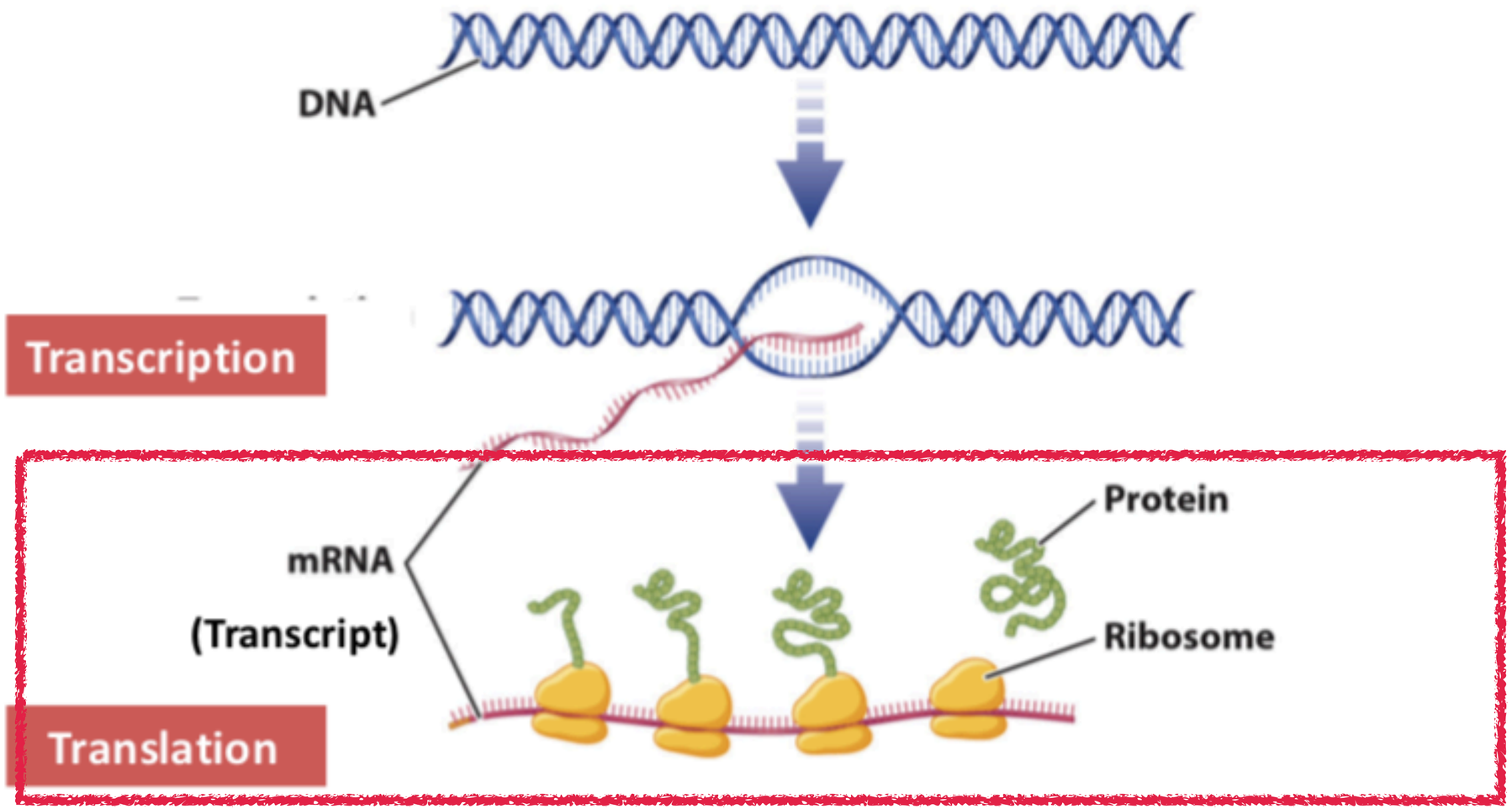
4 June 2021

Gene Regulation



from https://rbpaonline.com/

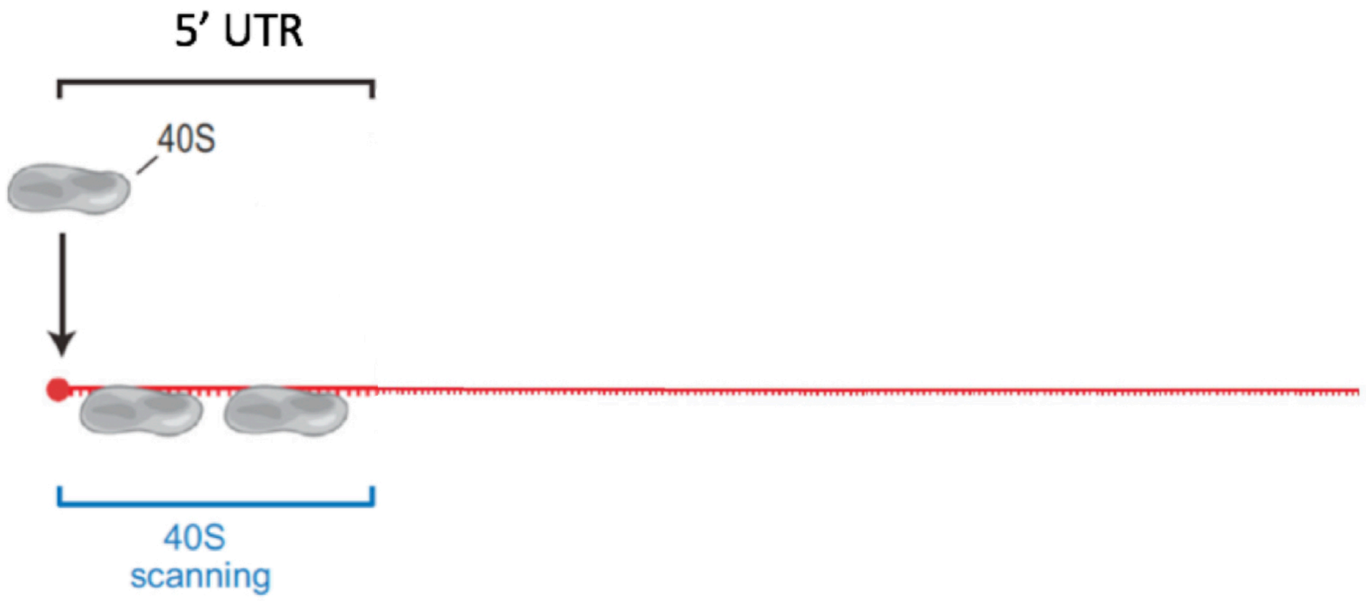
Gene Regulation



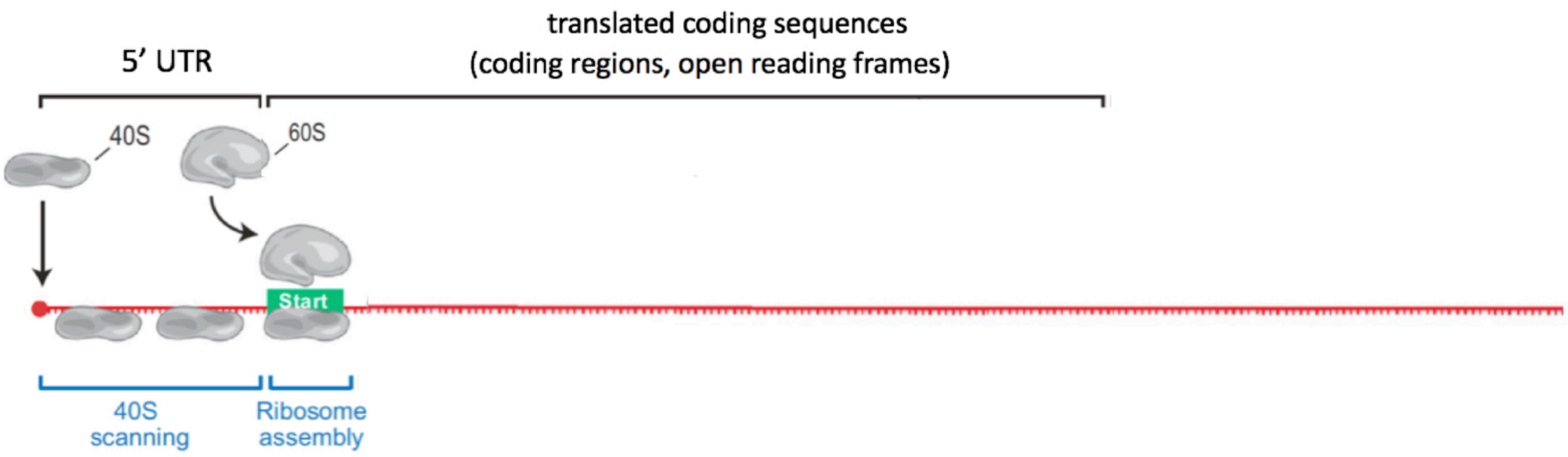
from https://rbpaonline.com/

Translation Mechanism

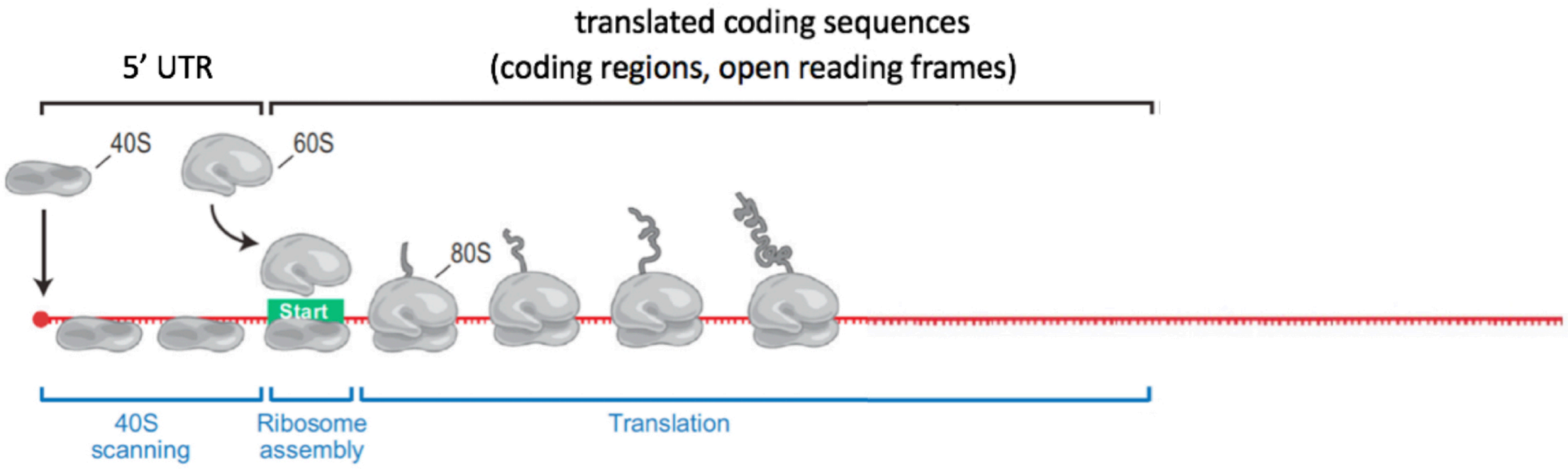
Translation Mechanism



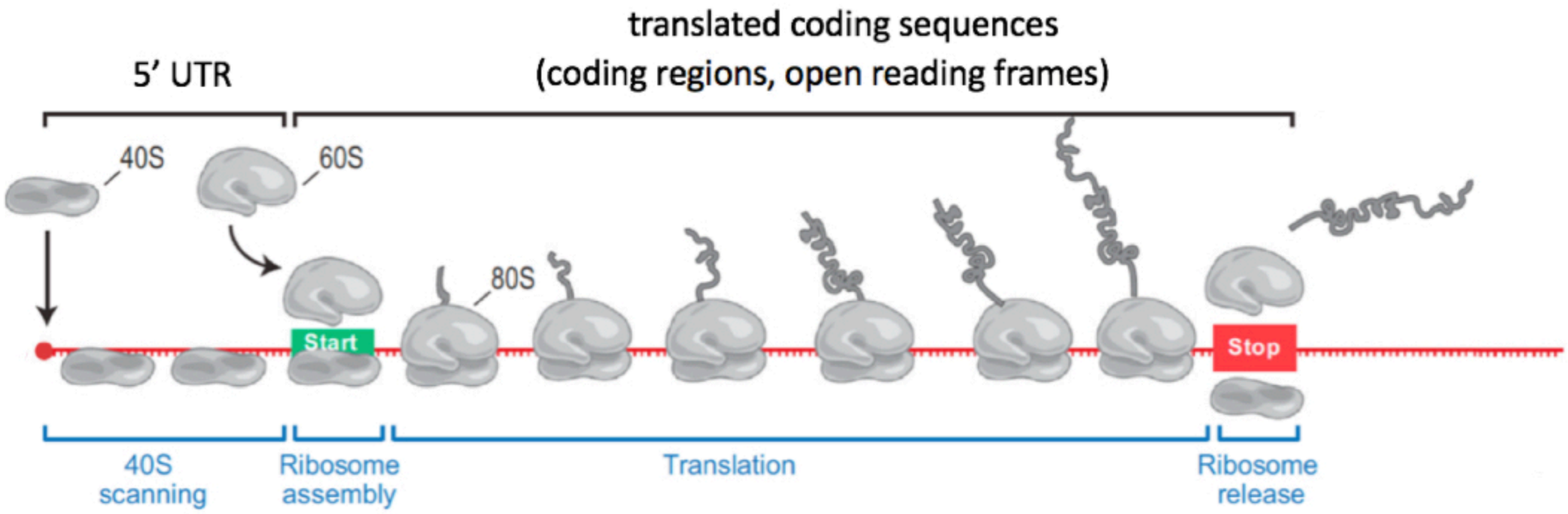
Translation Mechanism



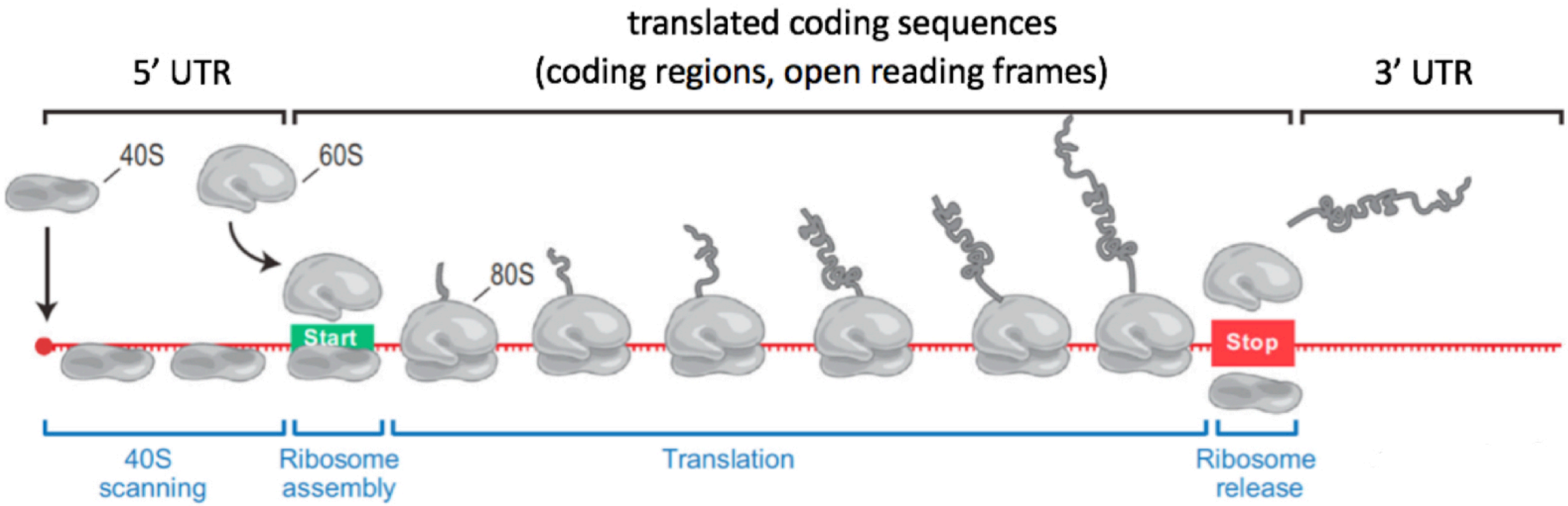
Translation Mechanism



Translation Mechanism



Translation Mechanism



Upstream Open Reading Frames (uORFs)

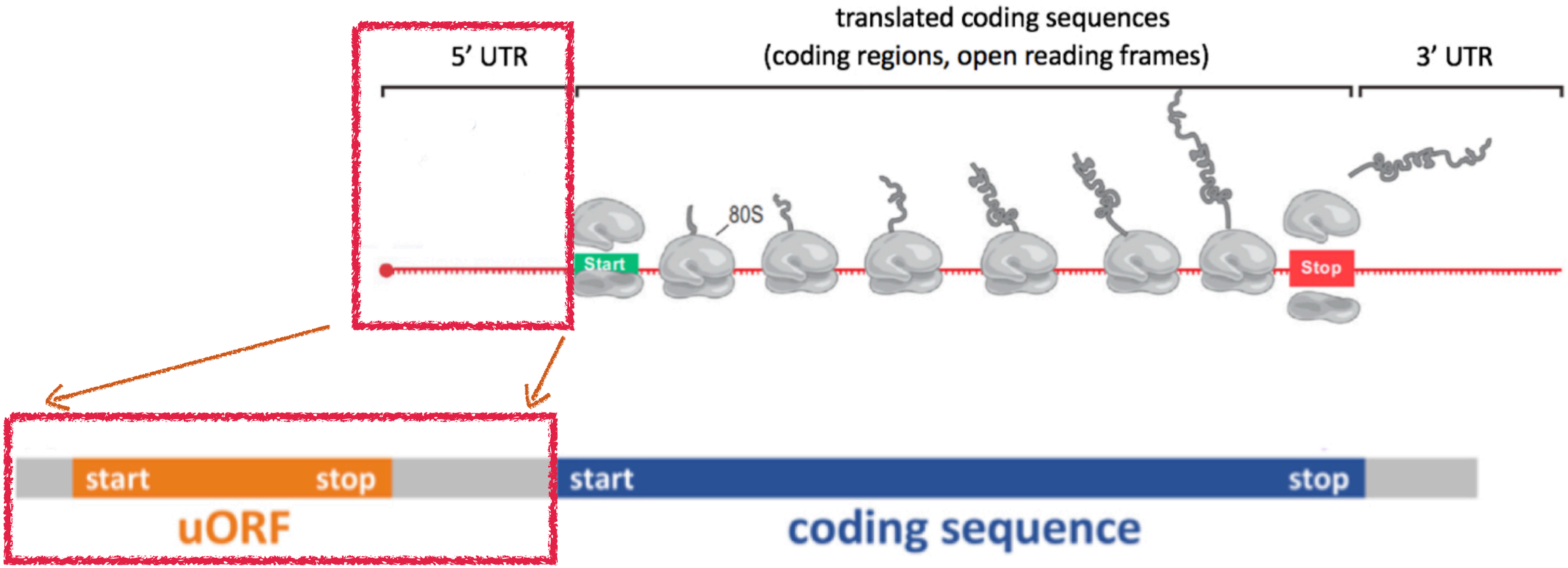
Pairs of start and stop codons located

upstream of protein coding sequences (5’UTR)

Upstream Open Reading Frames (uORFs)

Pairs of start and stop codons located

upstream of protein coding sequences (5'UTR)



Upstream Open Reading Frames (uORFs)

Translation at uORFs affects translation at downstream coding region

(Arcondéguy el at. 2013, Struck et al.2016)

However, the mechanism is **unclear**

Upstream Open Reading Frames (uORFs)

However, the mechanism is **unclear**
Because locations of translated uORFs are
NOT identified

Can we identify translated upstream Open Reading Frames (uORFs)?

Aim

Find locations of

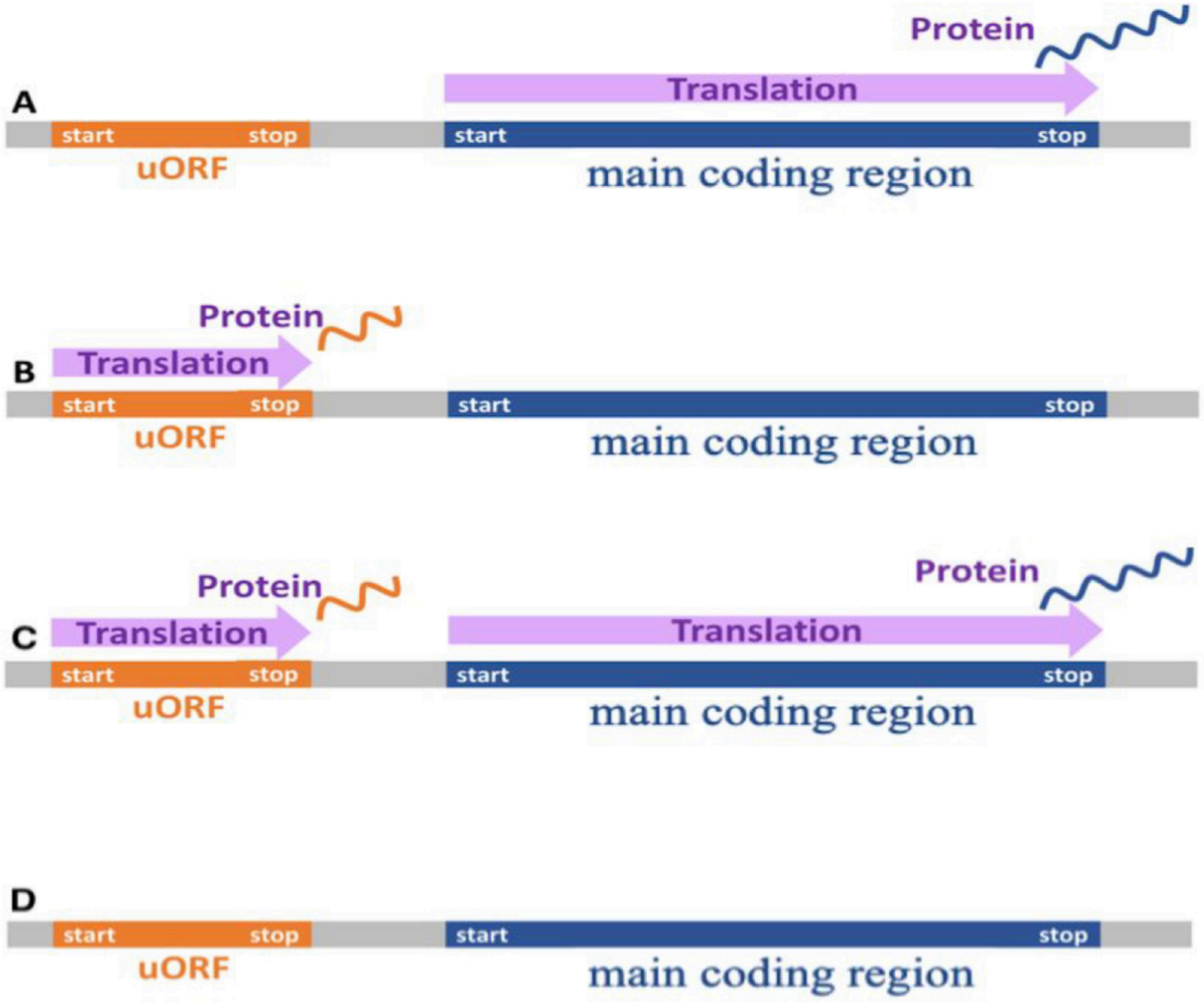
1. translated upstream Open Reading Frames (uORFs)

Aim

Find locations of

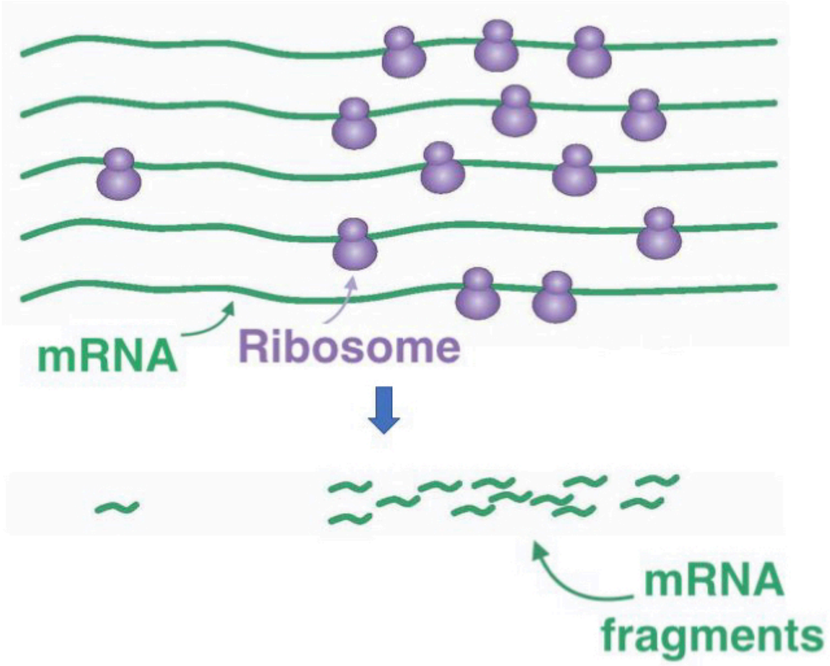
1. translated upstream Open Reading Frames (uORFs)
2. main coding regions

Cases


A
Protein
Translation
start
stop
start
stop
uORF
main coding region
B
Protein
Translation
start
stop
start
stop
uORF
main coding region
C
Protein
Translation
Protein
Translation
start
stop
start
stop
uORF
main coding region
D
start
stop
start
stop
uORF
main coding region

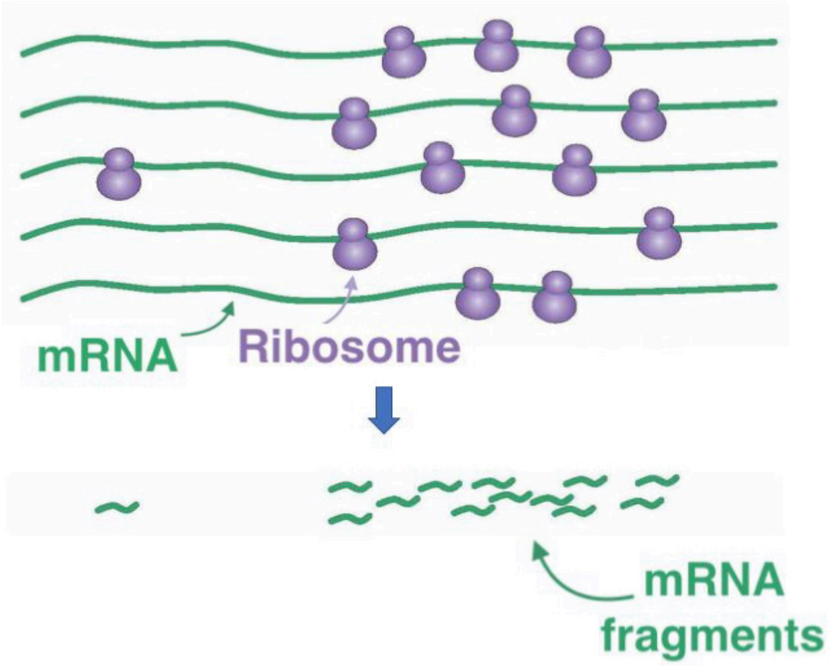
How?

Ribosome footprint



Isolate and sequence short mRNA fragments

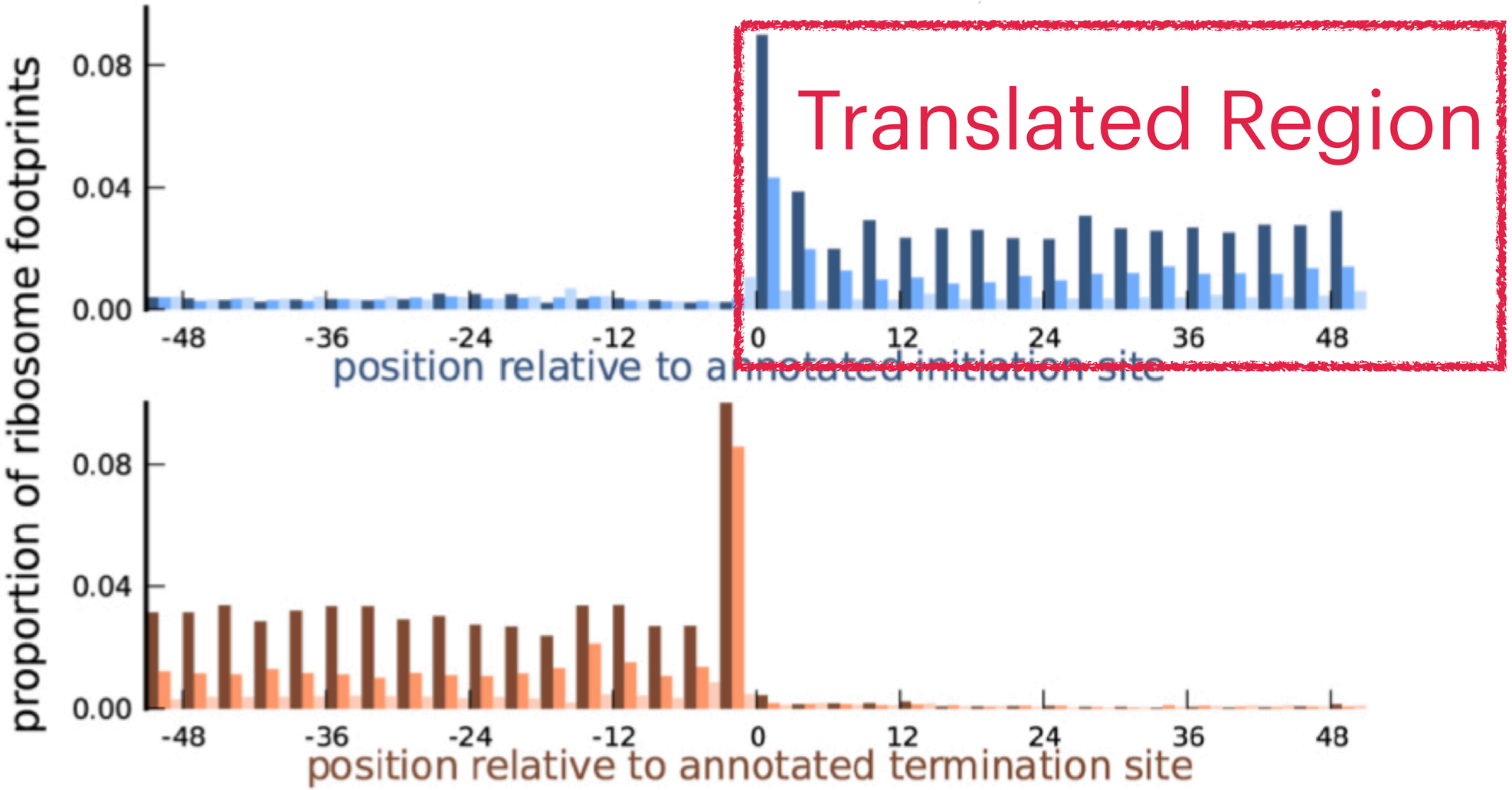
Ribosome footprint



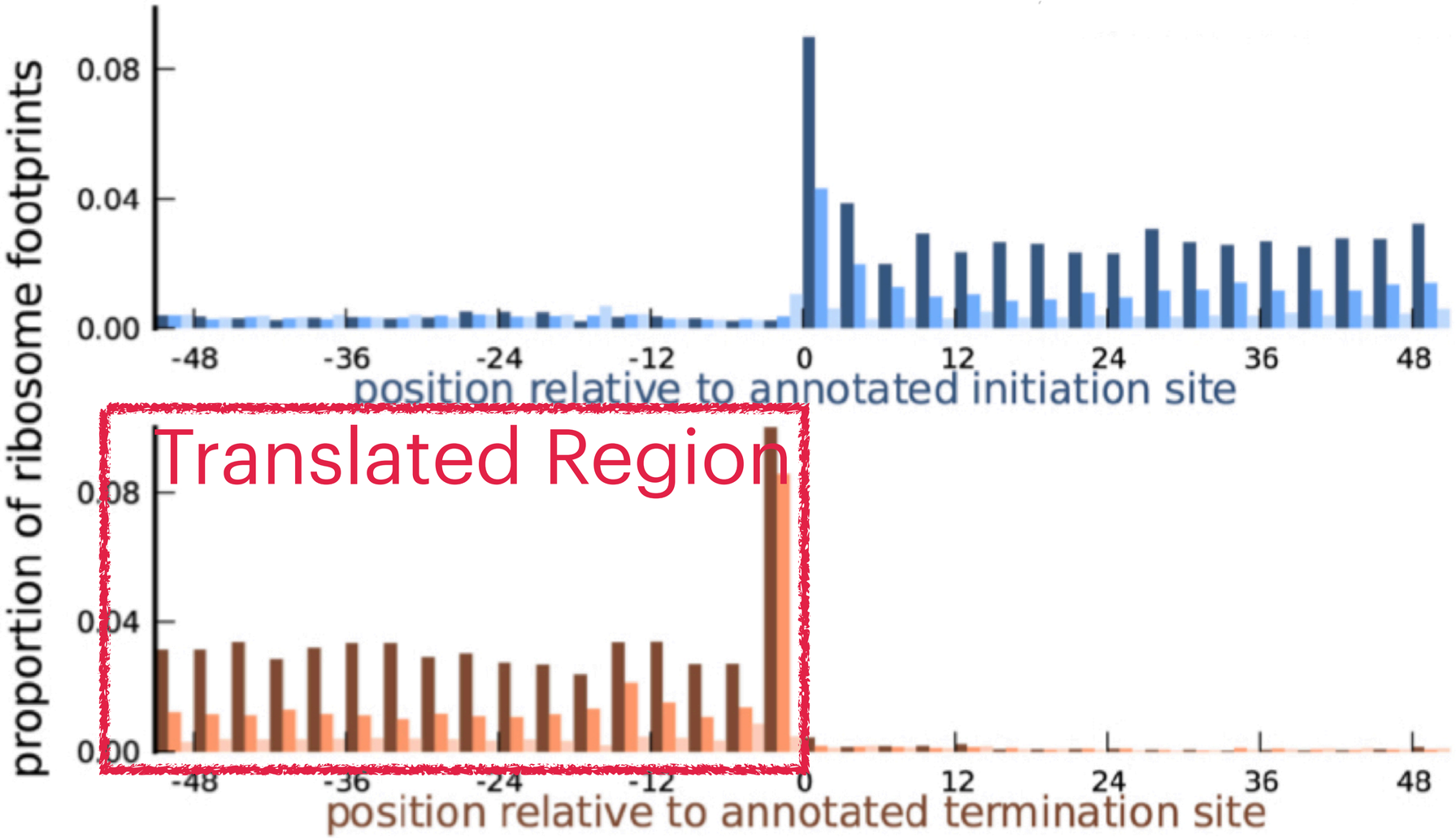
Isolate and sequence short mRNA fragments

Provide information about translated parts

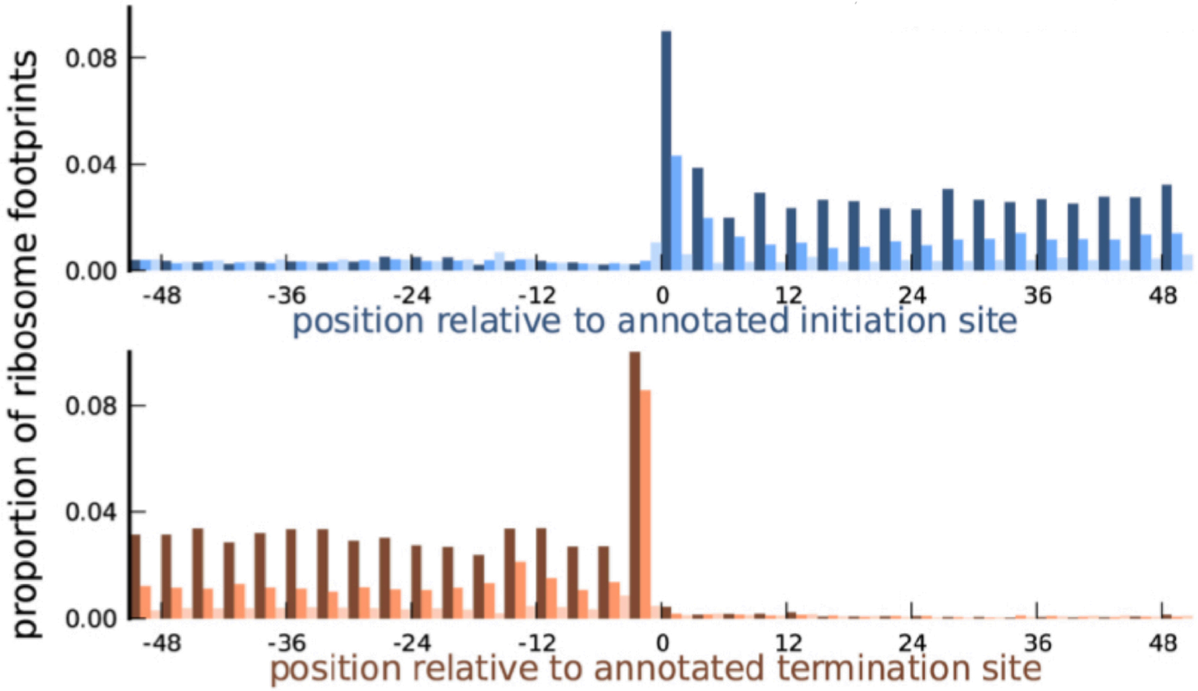
Ribosome footprint



Ribosome footprint

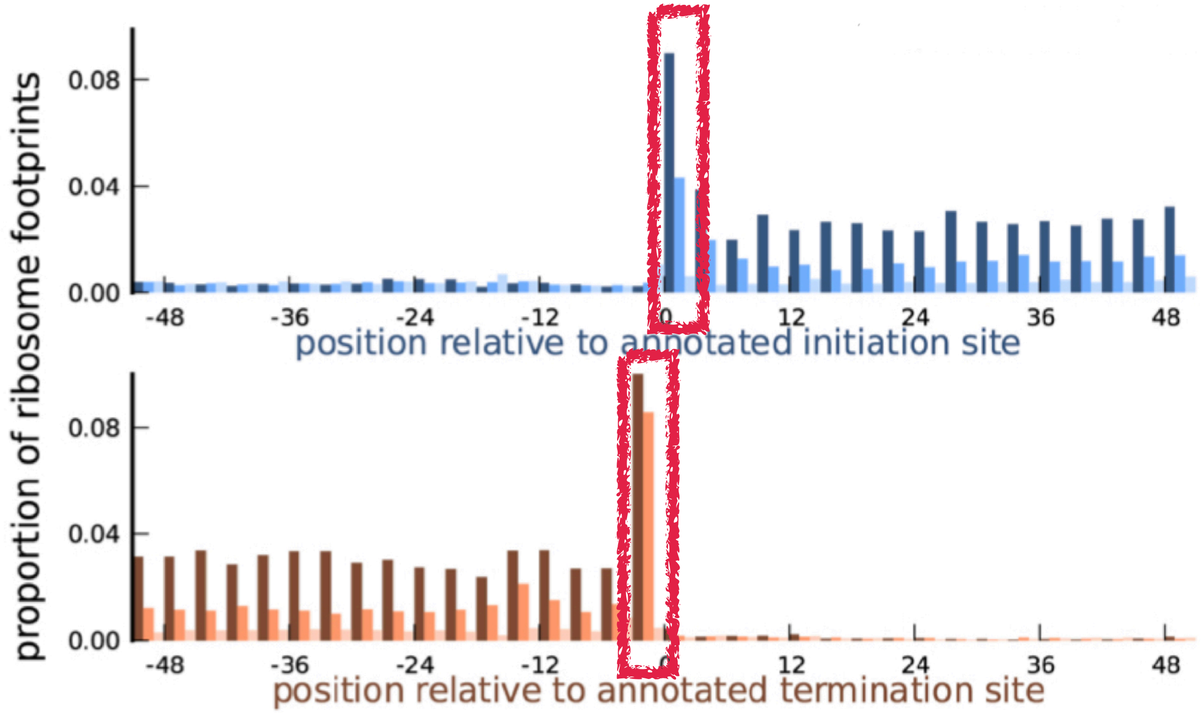


Ribosome footprint Characteristics



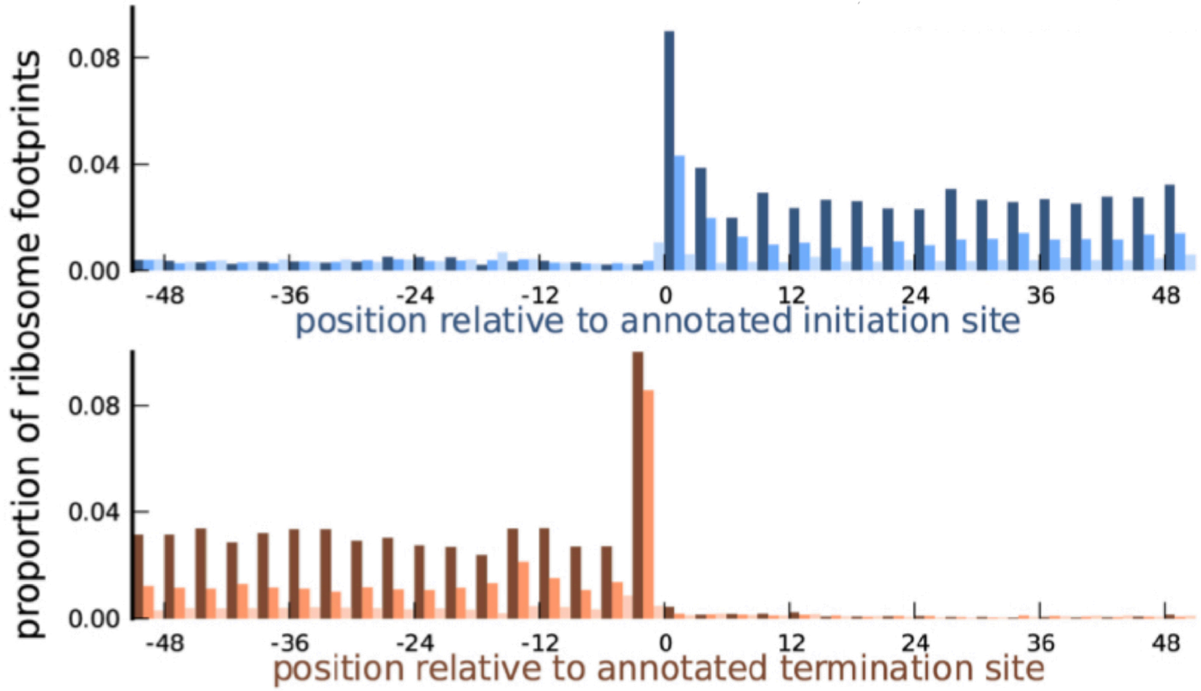
1. Higher counts within translated region

Ribosome footprint Characteristics



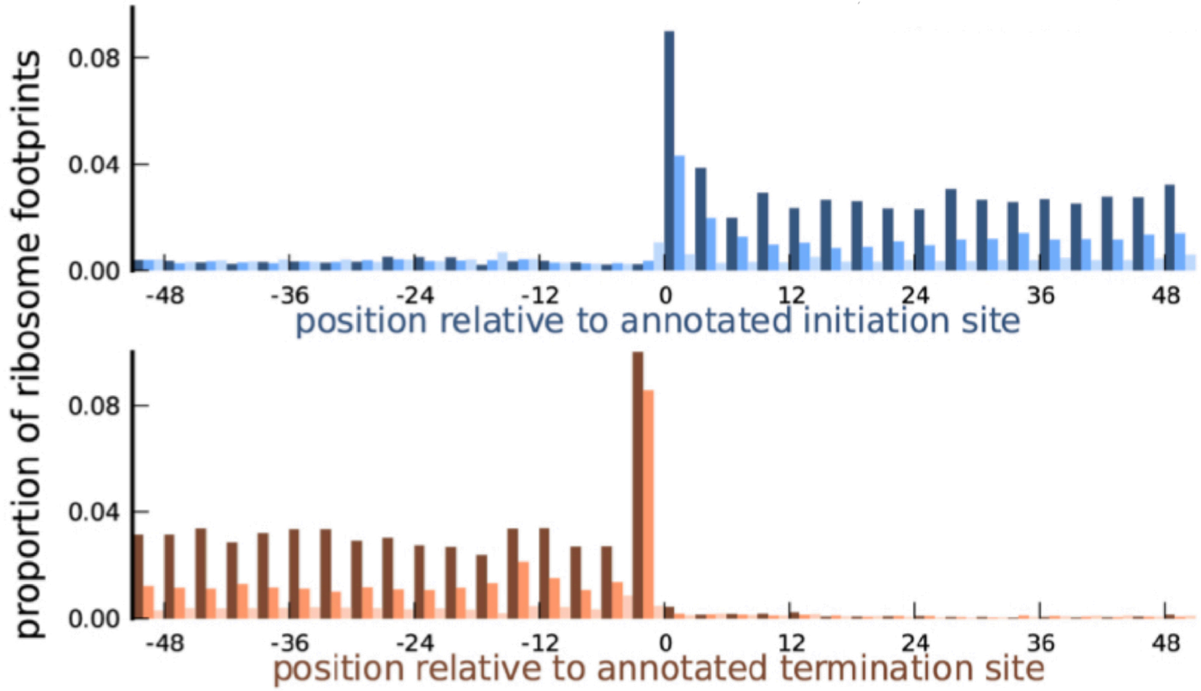
1. Higher counts within translated region
2. 3-base periodicity within translated region

Ribosome footprint Characteristics



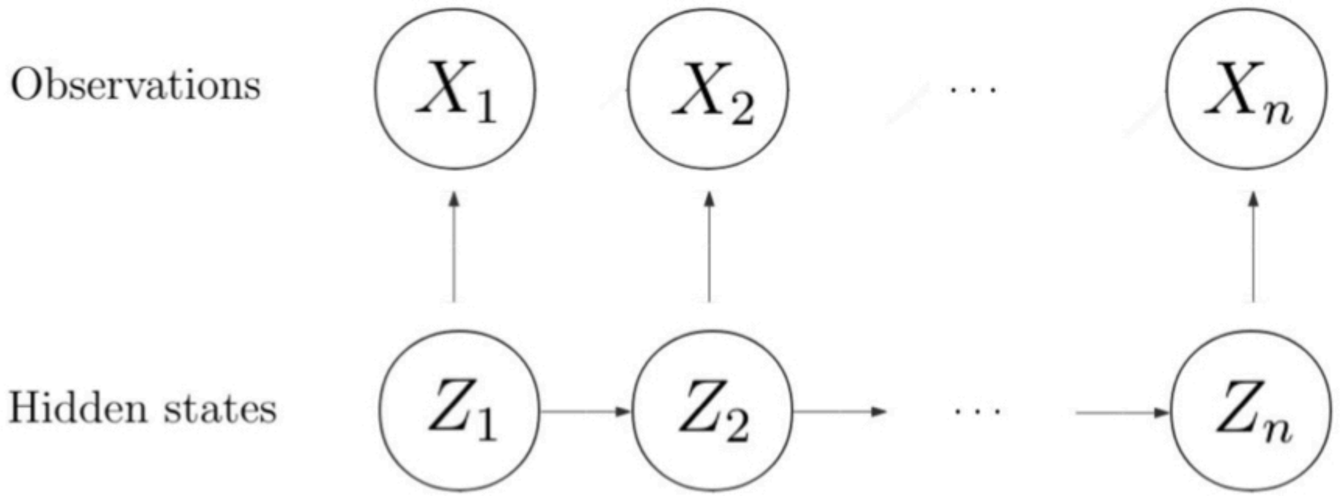
1. Higher counts within translated region
2. 3-base periodicity within translated region
3. Different pattern at translation start and stop sites

Ribosome footprint Characteristics

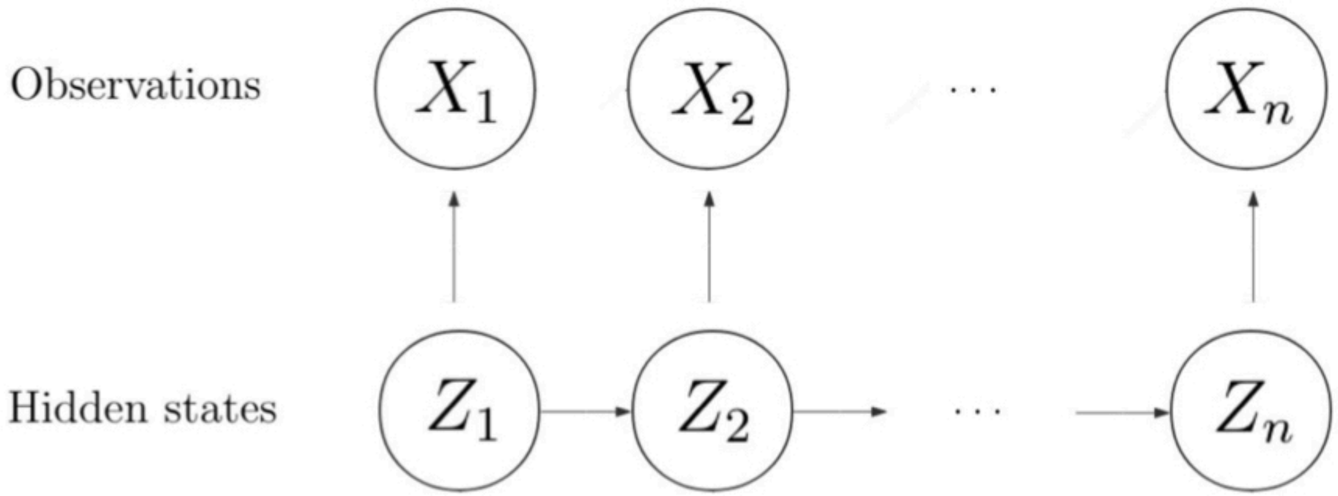


Counts depends on different states

Hidden Markov Model

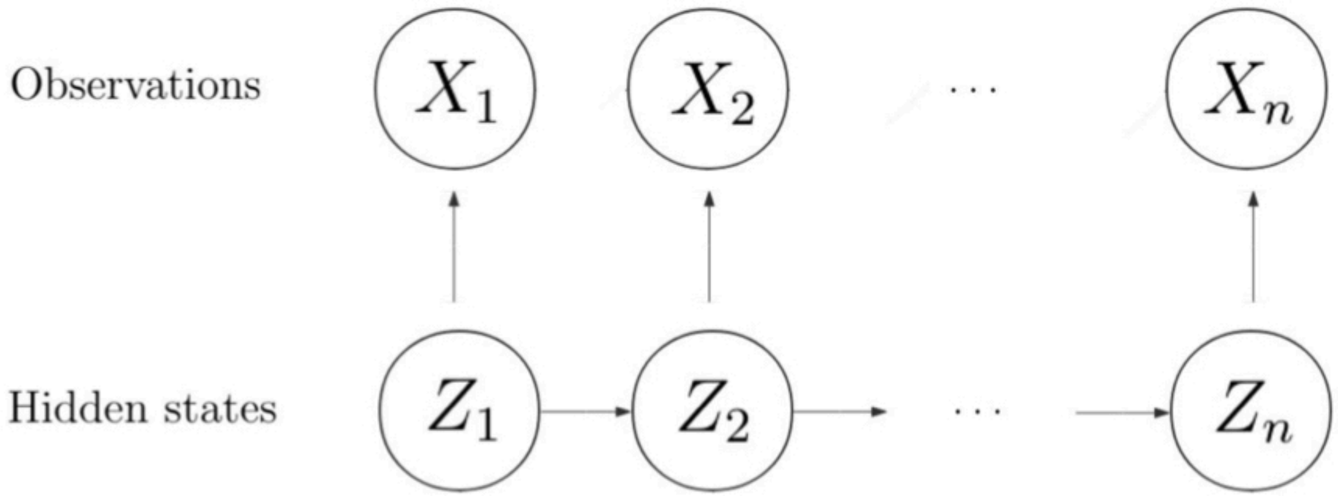


Hidden Markov Model



Observations: ribosome footprint counts

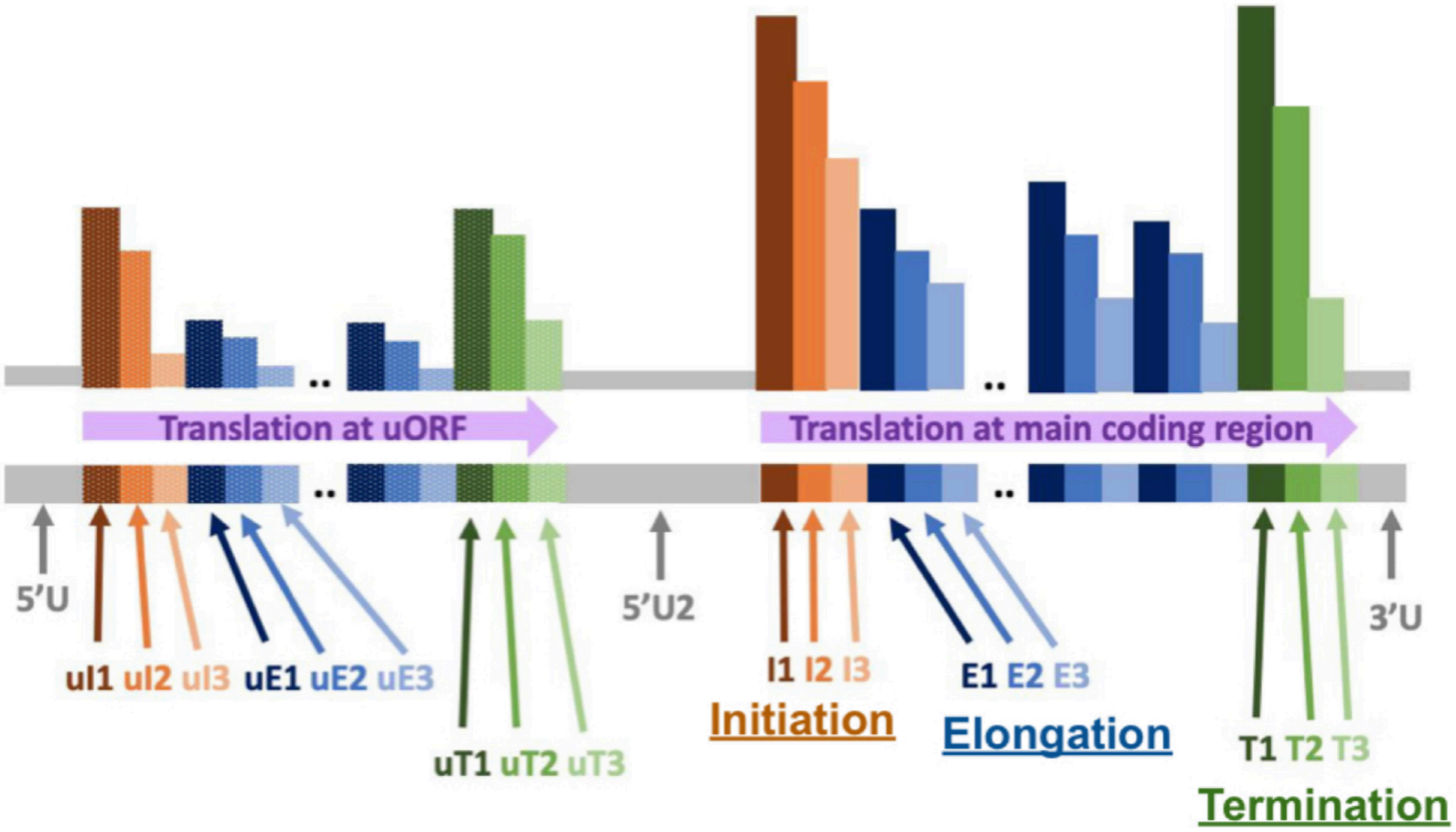
Hidden Markov Model



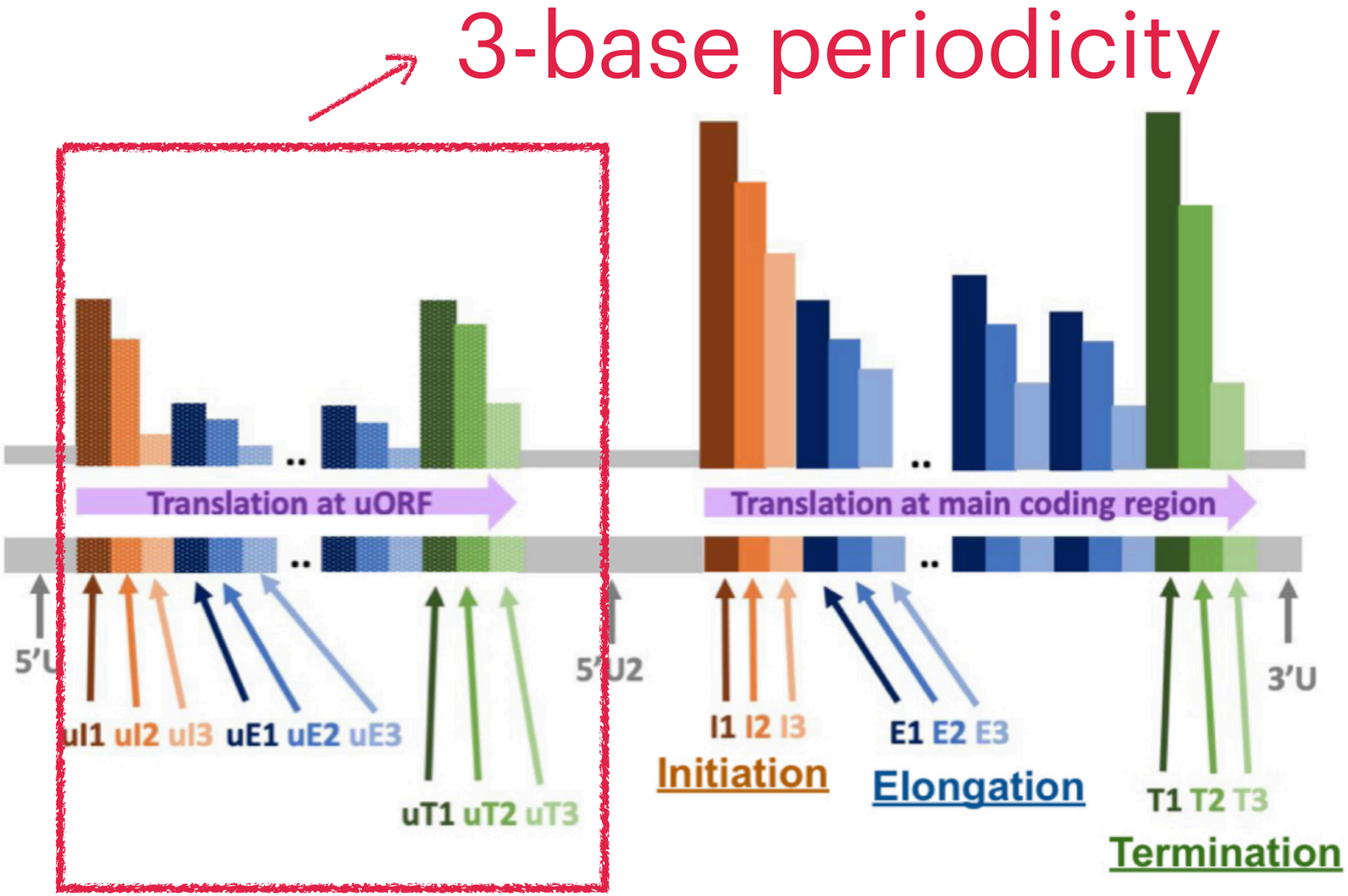
Observations: ribosome footprint counts

Hidden states: 21 translational states

21 Hidden States

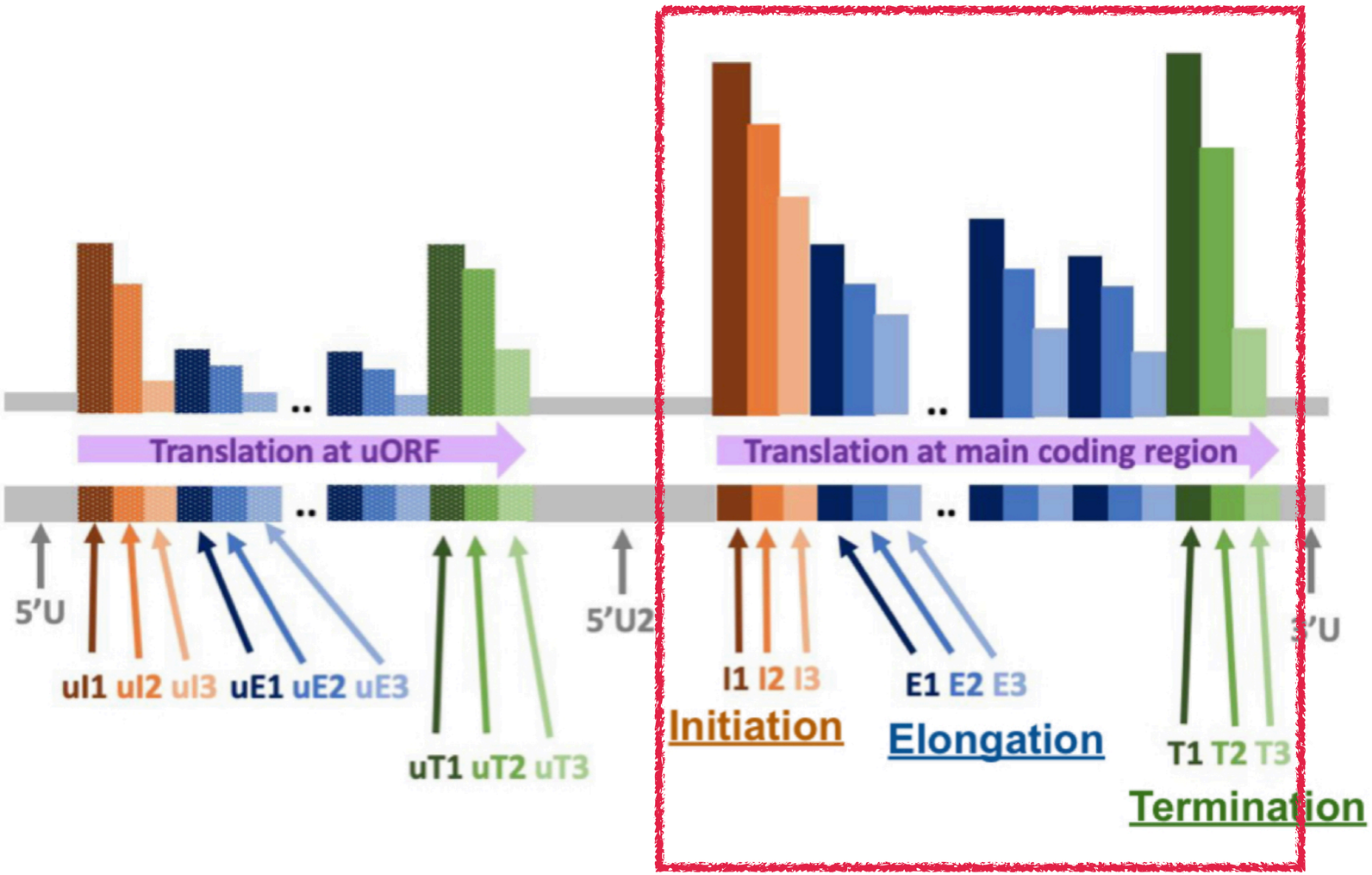


21 Hidden States

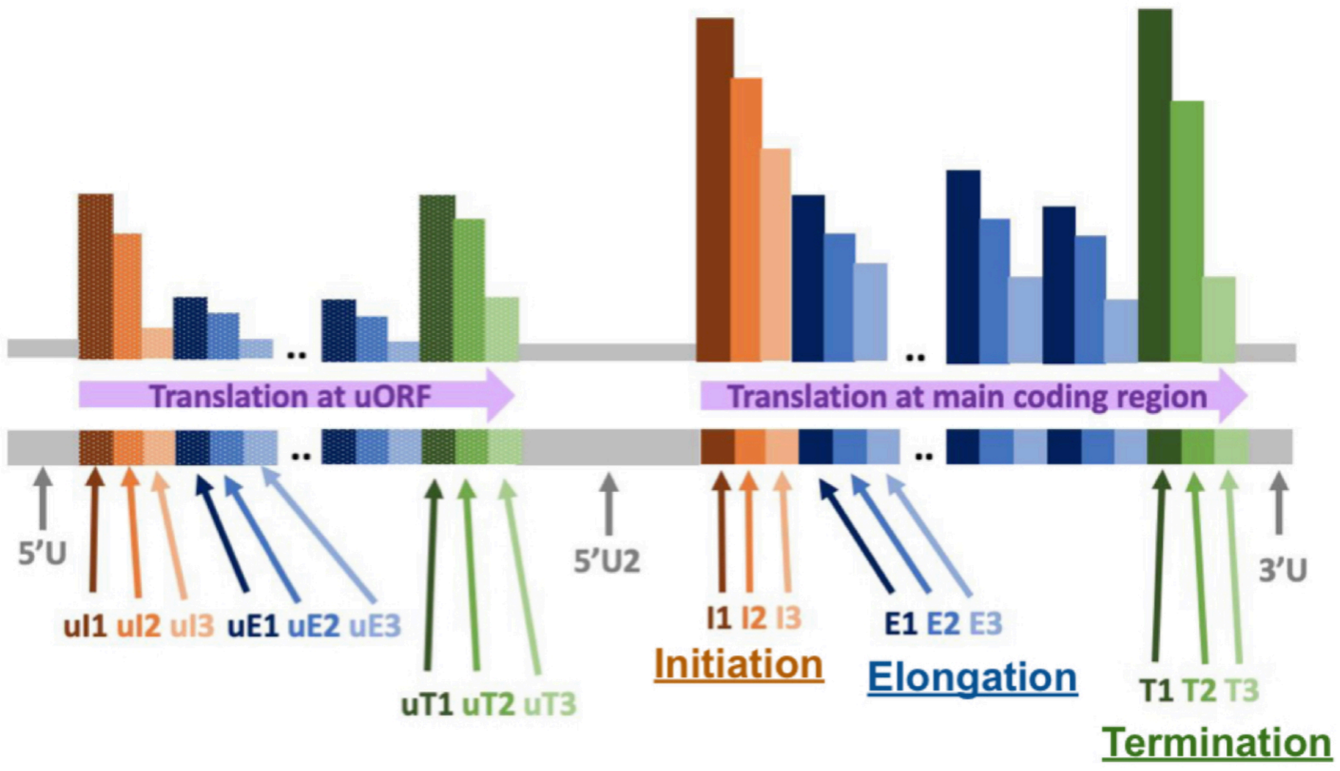


21 Hidden States

3-base periodicity

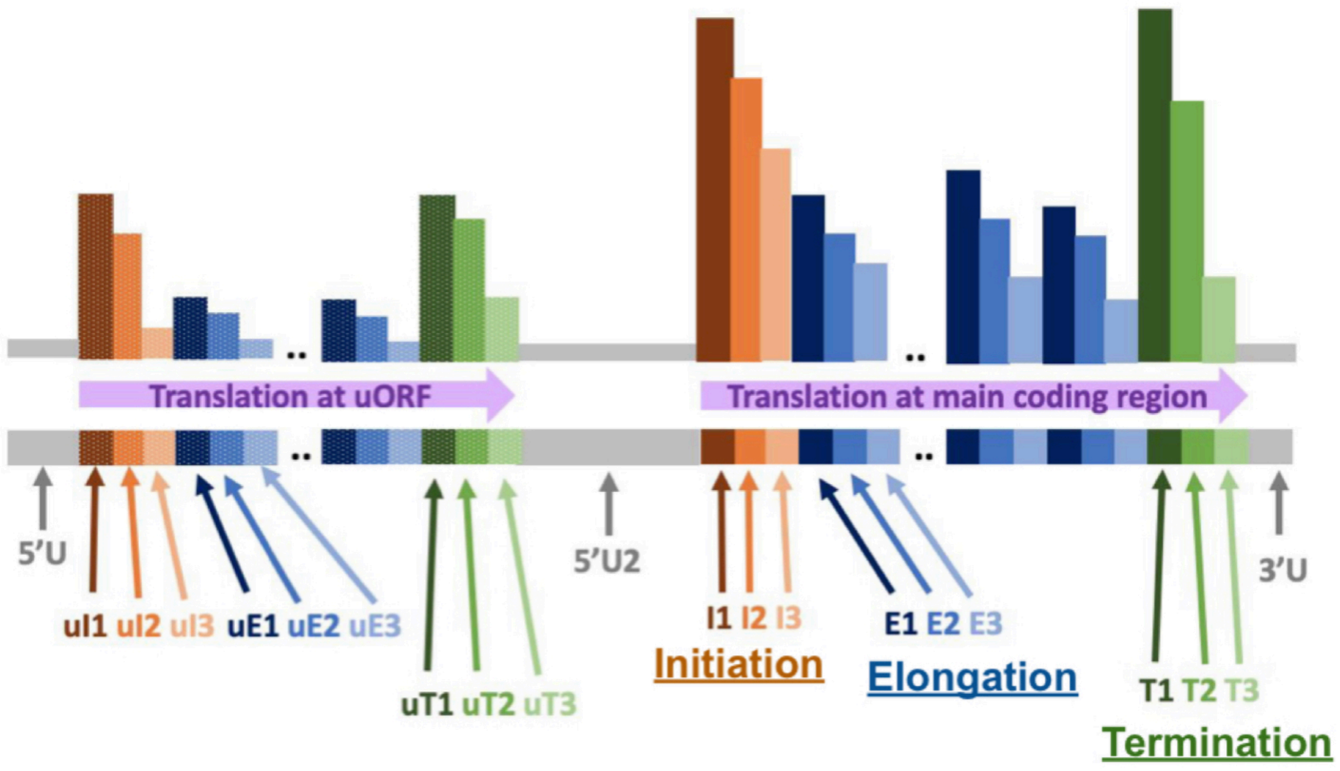


Assumptions



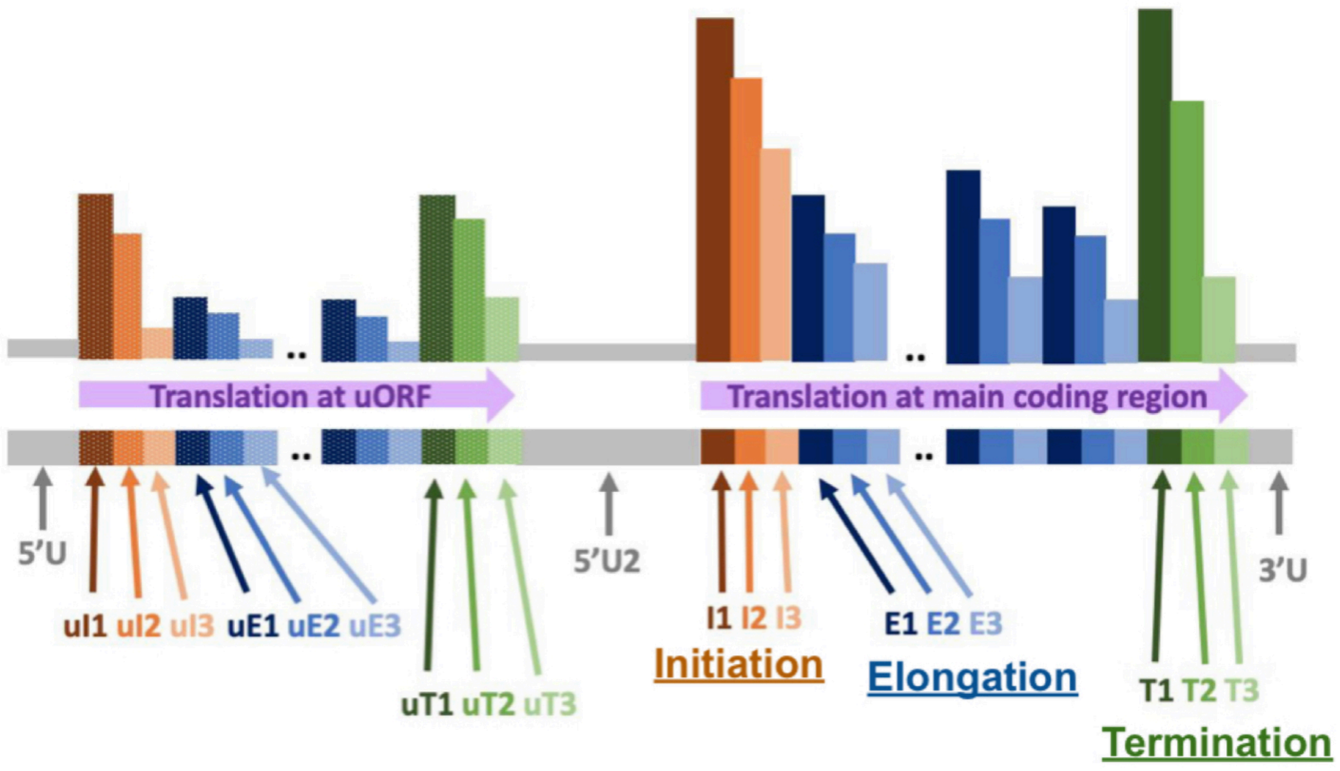
1. Starts from 5'U

Assumptions



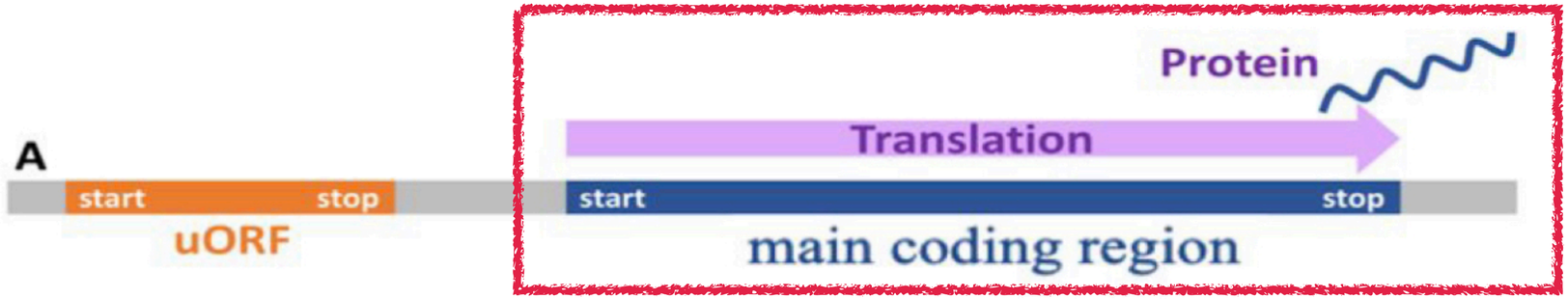
1. Starts from 5'U
2. Transitional initial can occur at **non-canonical** start codon

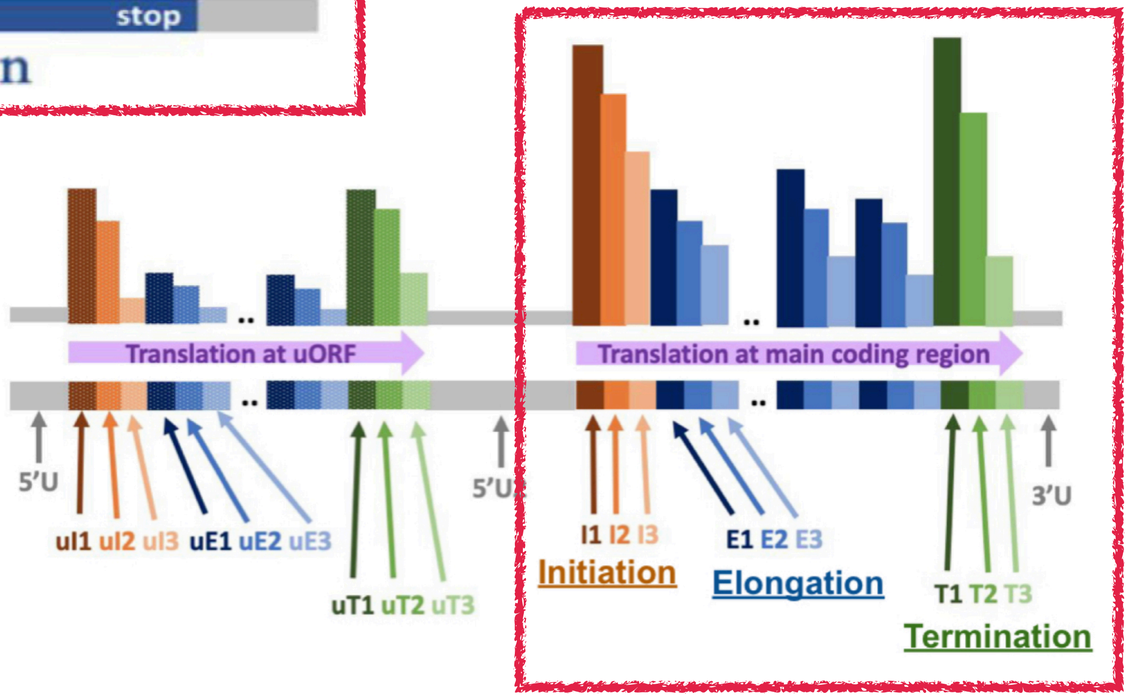
Assumptions



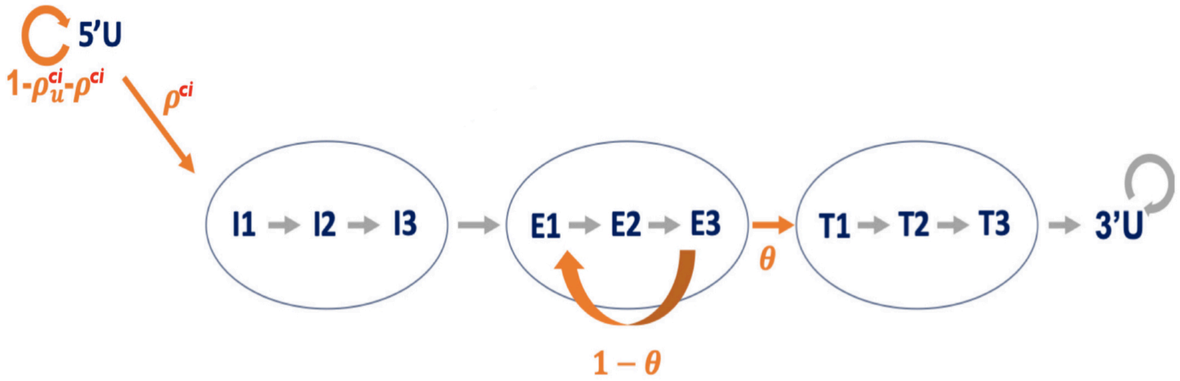
1. Starts from 5'U
2. Transitional initial can occur at non-canonical start codon
3. Transitional termination occurs at first canonical stop codon
    - UAA, UGA, UAG

Transition probabilities — **Only main coding region**

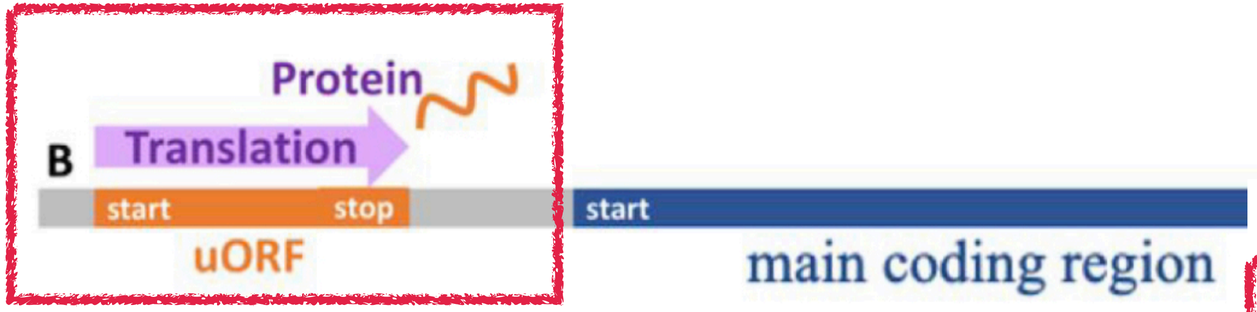


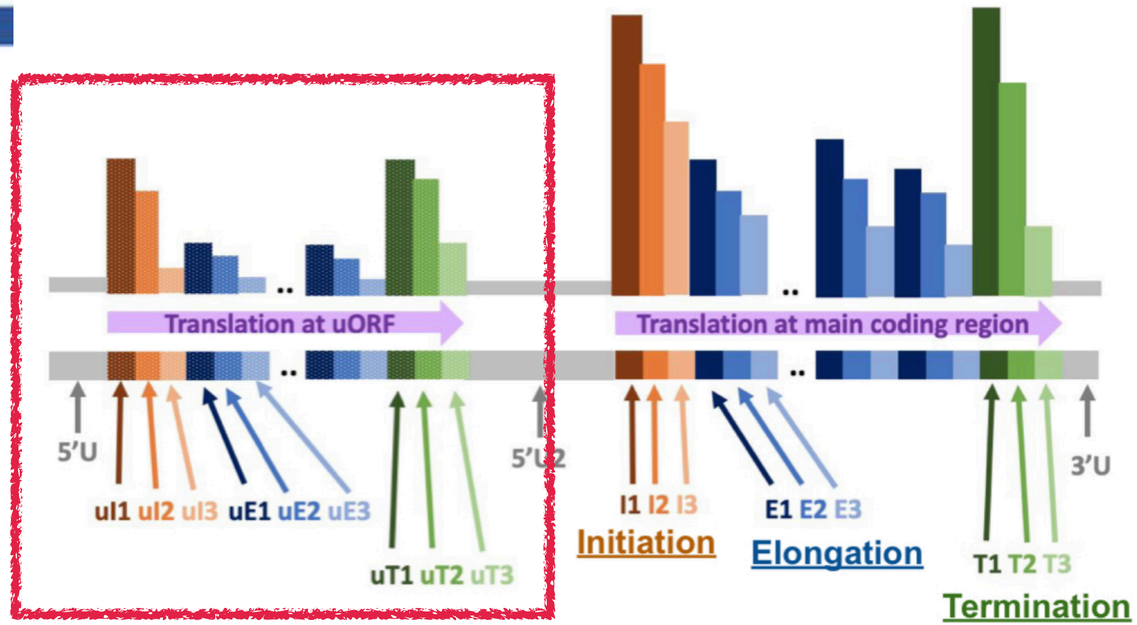


For **specific** start codon **ci**:

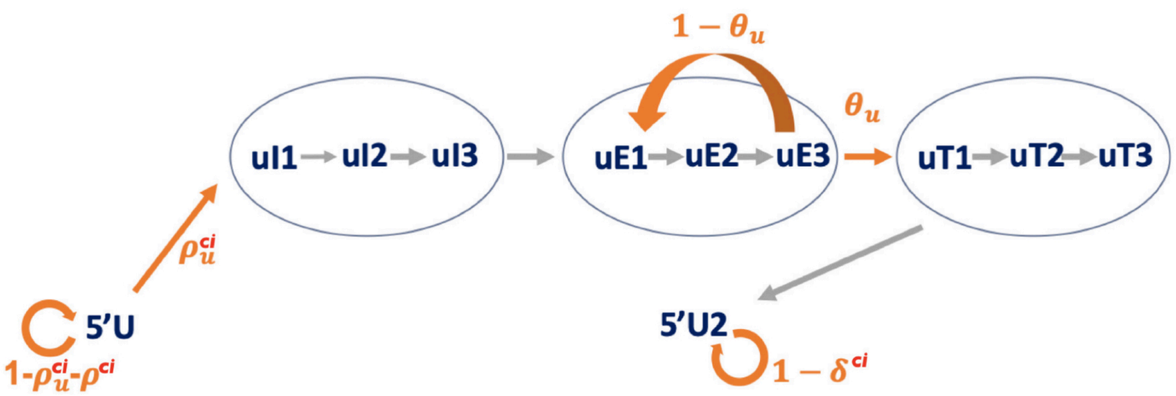


Transition probabilities — **Only uORF**



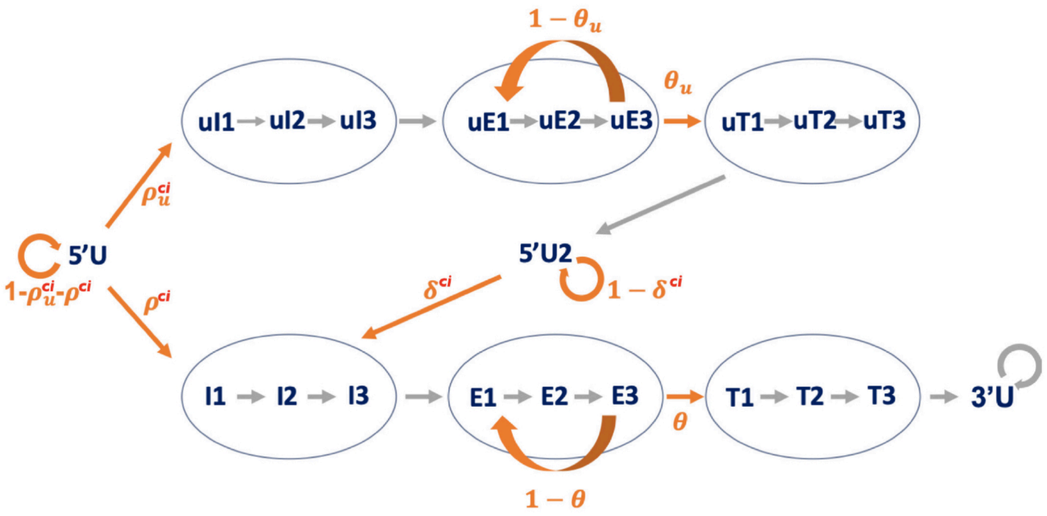


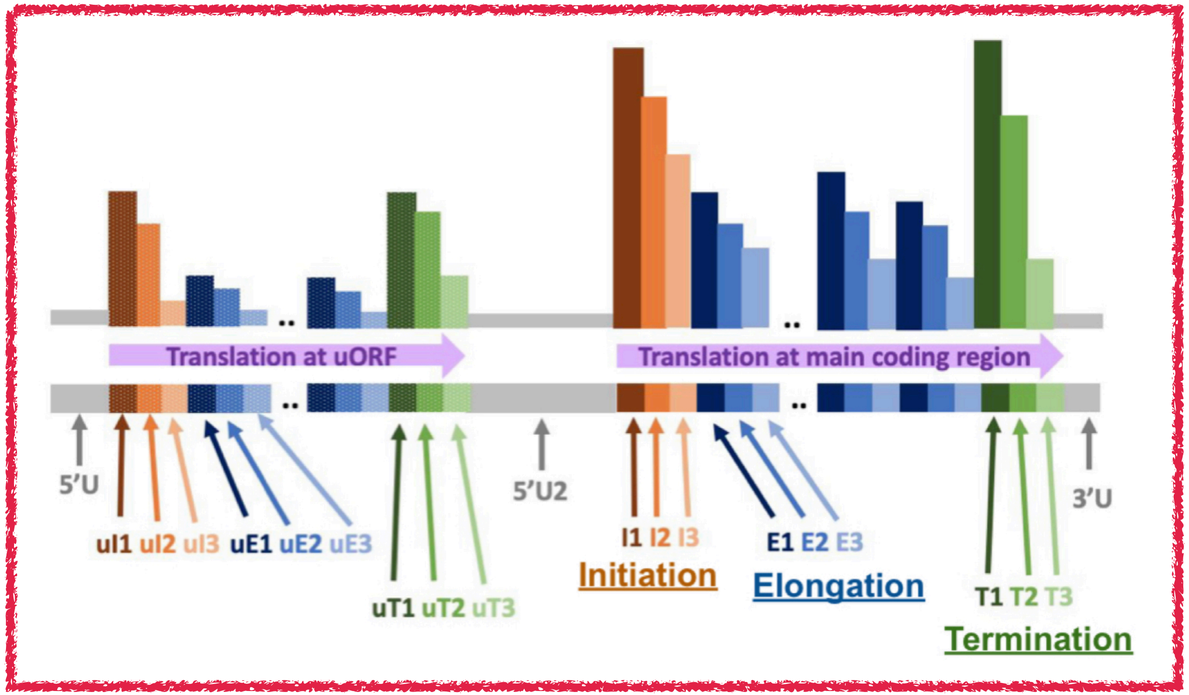
For **specific** start codon **ci**:



Transition probabilities — **uORF & main coding region**

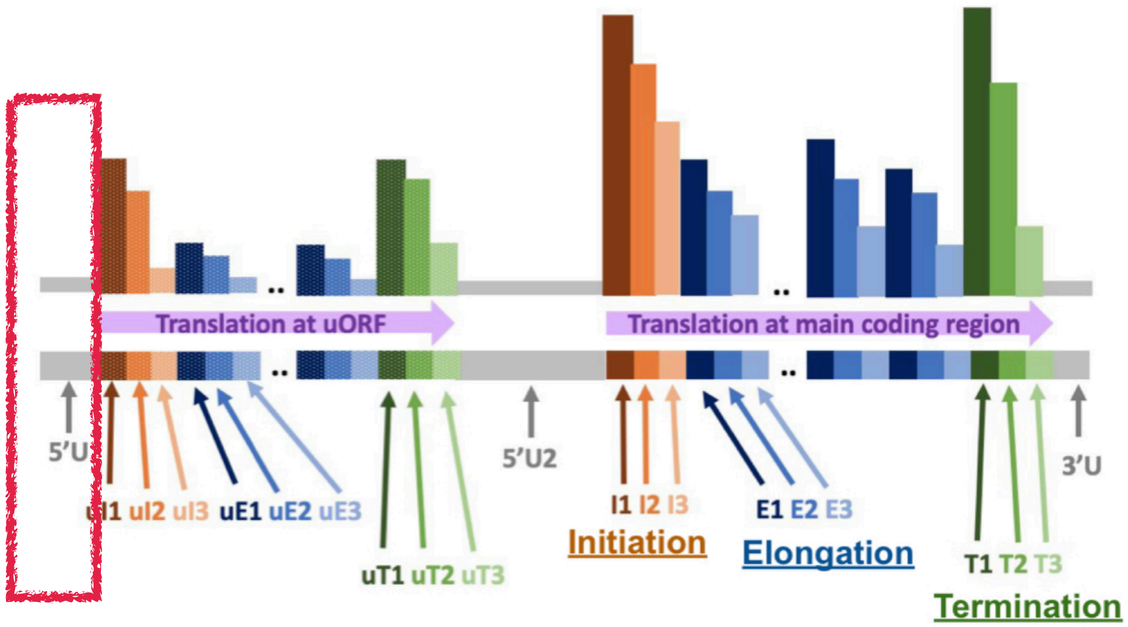
For **specific** start codon **ci**:



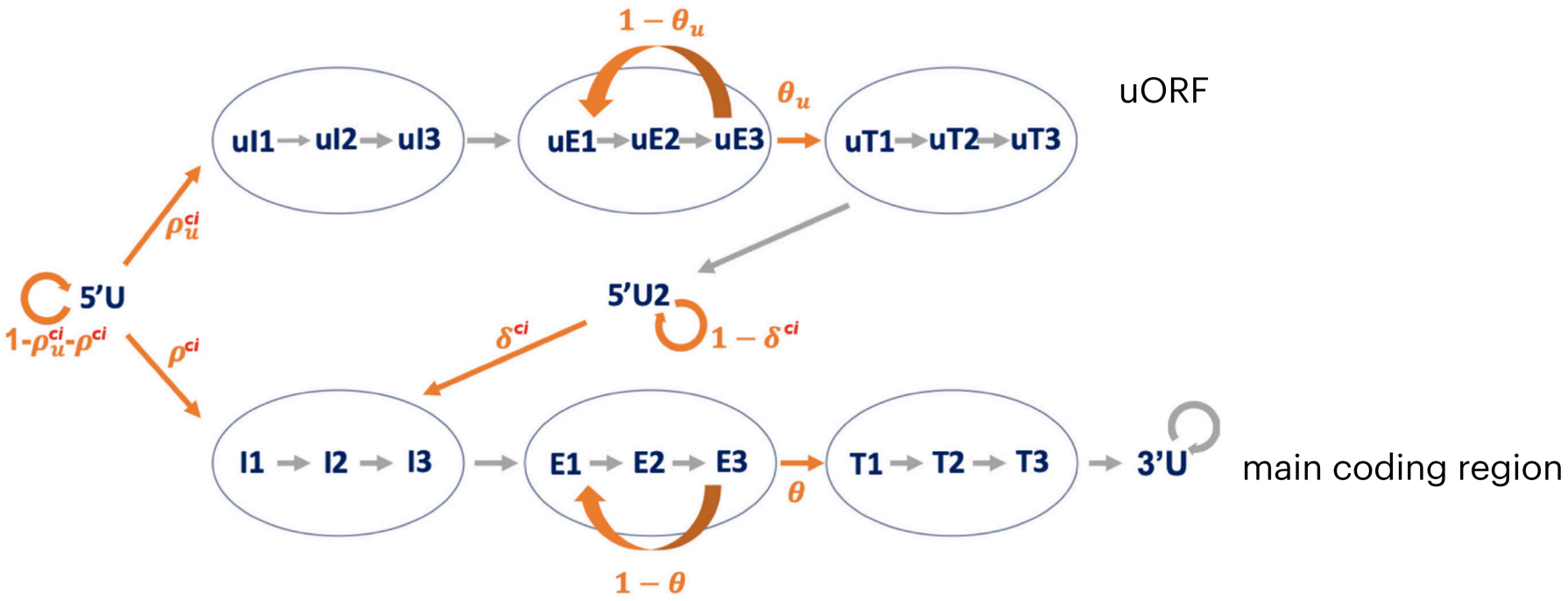


Transition probabilities — **None**

For **specific** start codon **ci**:



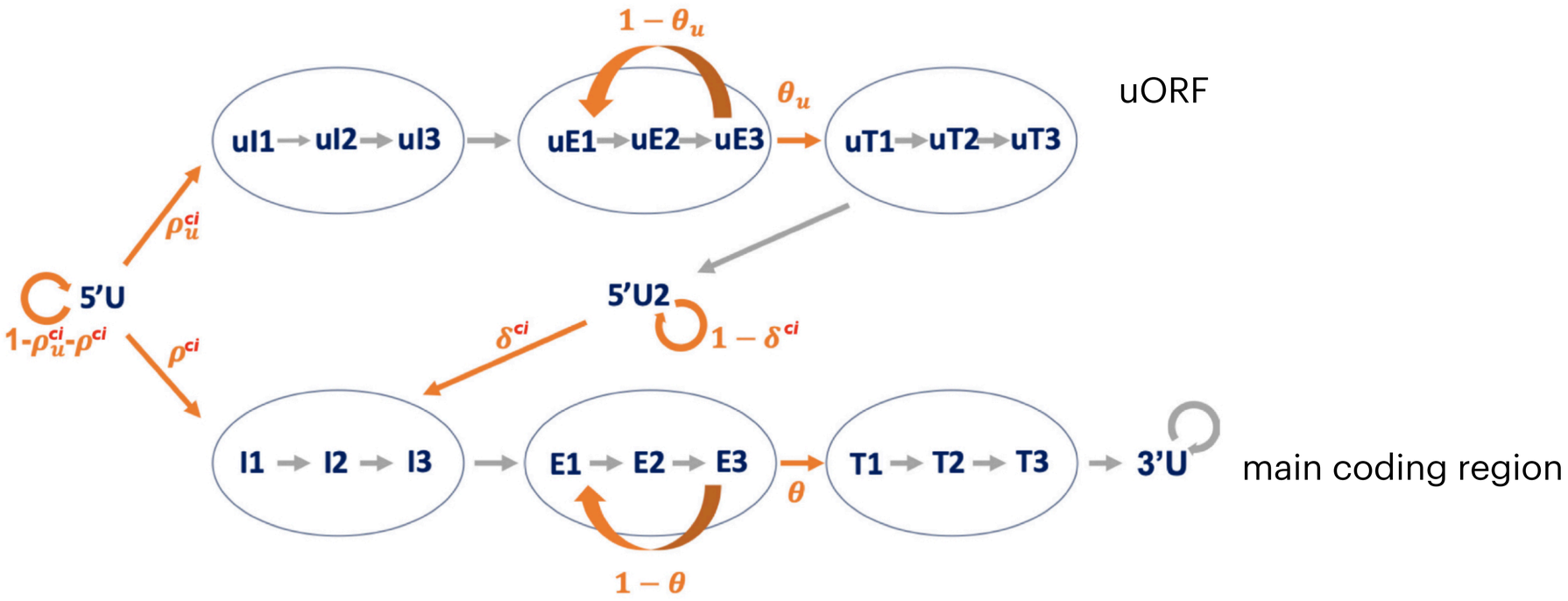
Transition probabilities — **Parameters**



For specific start codon ci:

- ρ_μ^{ci}
- ρ^{ci}
- δ^{ci}

Transition probabilities — **Parameters**



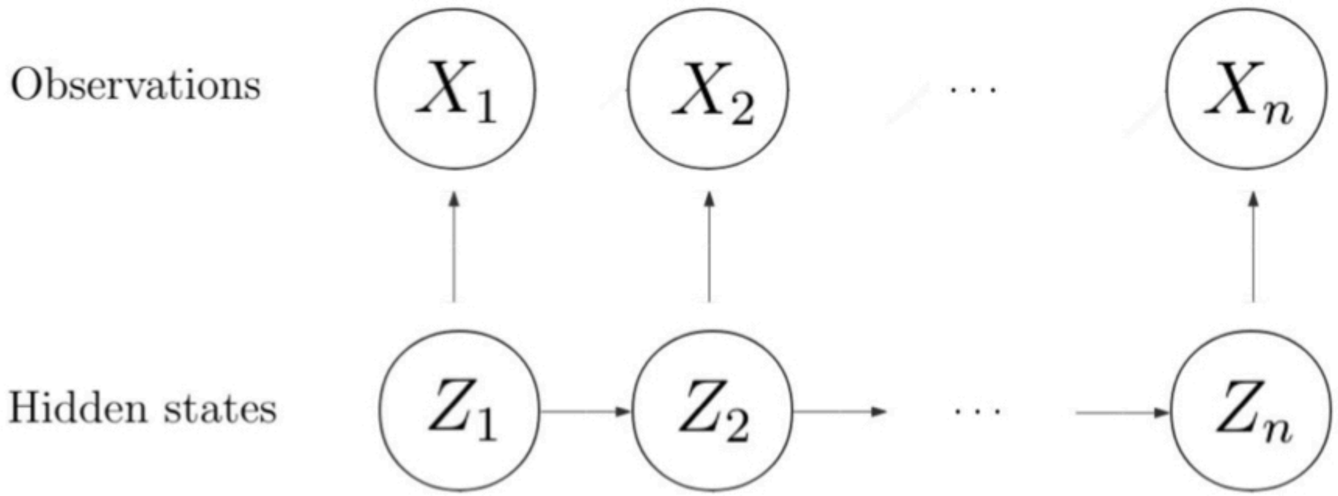
For specific start codon ci:

- ρ_μ^{ci}
- ρ^{ci}
- δ^{ci}

For stop codon:

- θ_μ
- θ

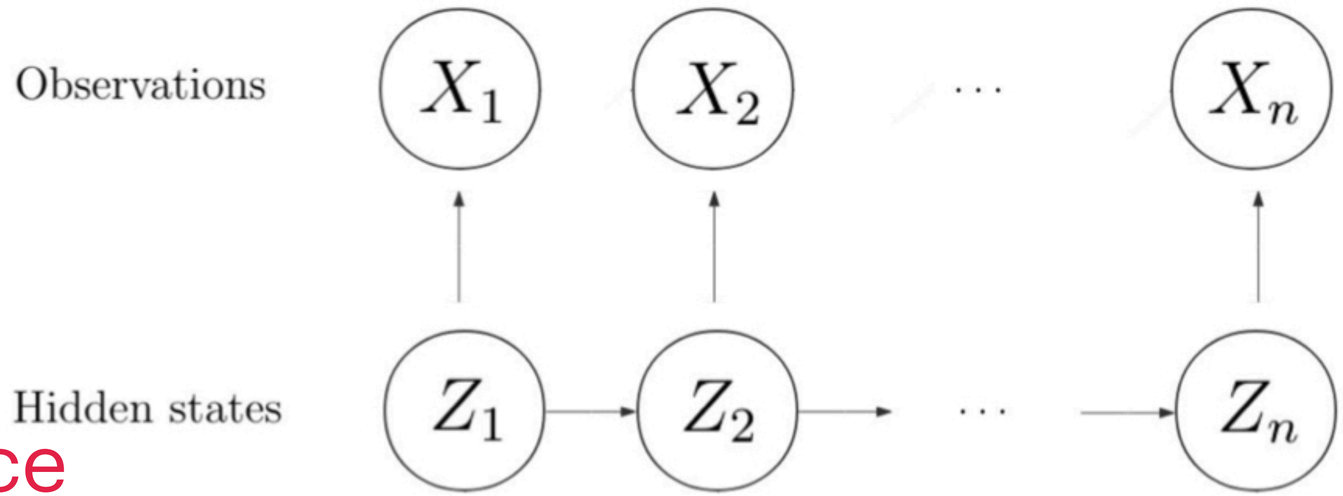
Emission probabilities — One RNA sequence



$$X_n \mid Z_n = z \sim \text{Poisson}(\boldsymbol{\lambda}_{z,n})$$

$$\boldsymbol{\lambda}_{z,\,n} \sim \text{Gamma}(\boldsymbol{\alpha}_z,\ \boldsymbol{\beta}_z)$$

Emission probabilities — Multiple (T) RNA sequences

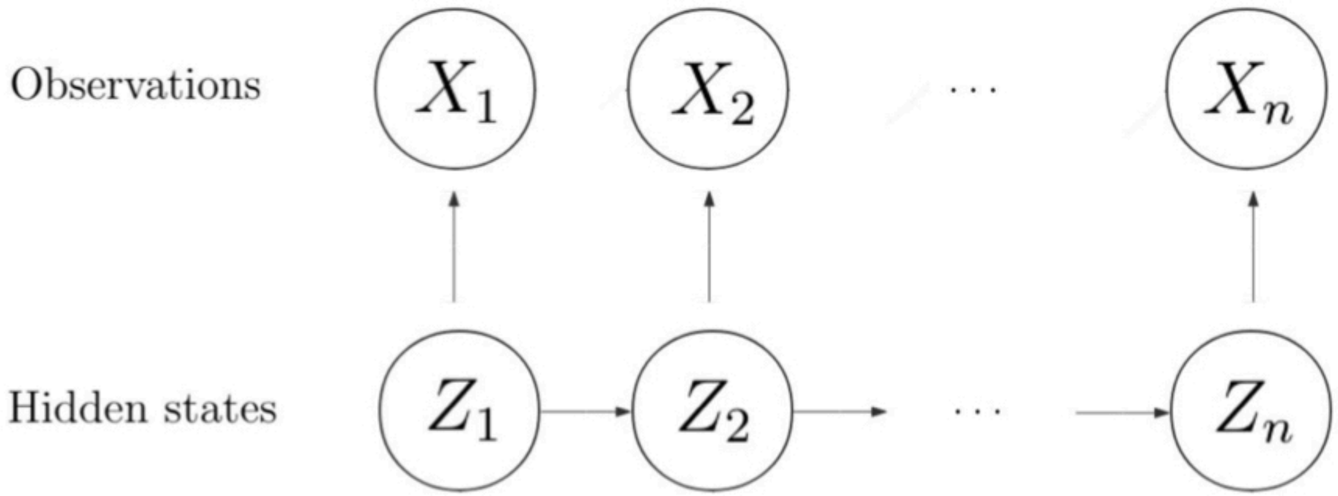


t^{th} sequence

$$X_n^{(t)} \mid Z_n^{(t)} = z \sim \text{Poisson}(\boldsymbol{\lambda}_{z,n}^{(t)})$$

$$\boldsymbol{\lambda}_{z,\,n}^{(t)} \sim \text{Gamma}(\boldsymbol{\alpha}_z, \boldsymbol{\beta}_z)$$

Emission probabilities — Multiple (T) RNA sequences

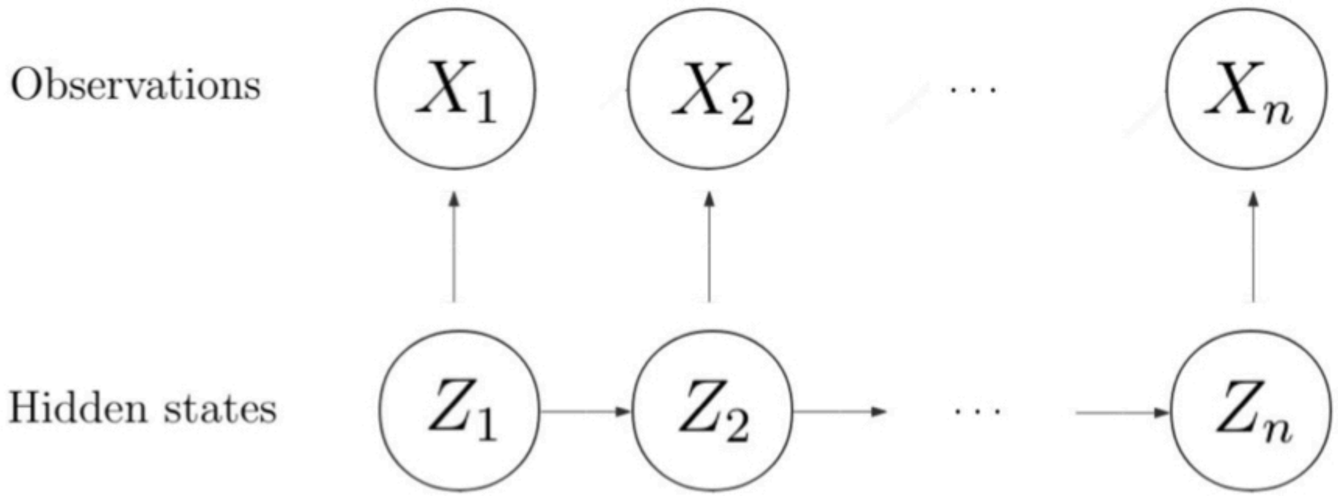


$$X_n^{(t)} \mid Z_n^{(t)} = z \sim \text{Poisson}(\boldsymbol{\lambda}_{z,n}^{(t)} E^{(t)})$$

$$\boldsymbol{\lambda}_{z,\,n}^{(t)} \sim \text{Gamma}(\boldsymbol{\alpha}_z, \boldsymbol{\beta}_z)$$

Normalization Factor

Emission probabilities — Multiple (T) RNA sequences



$$X_{z,n}^{(t)} \sim \text{Negative Binomial}(\boldsymbol{\alpha}_z, \boldsymbol{\beta}_z / (E^{(t)} + \boldsymbol{\beta}_z))$$

Derivation

Estimation for parameters: EM algorithm
Inference of hidden states: Viterbi Algorithm

Simulation Results

Simulate 30 datasets, each contains 100 RNA sequences

Simulation Results

Simulate 30 datasets, each contains 100 RNA sequences

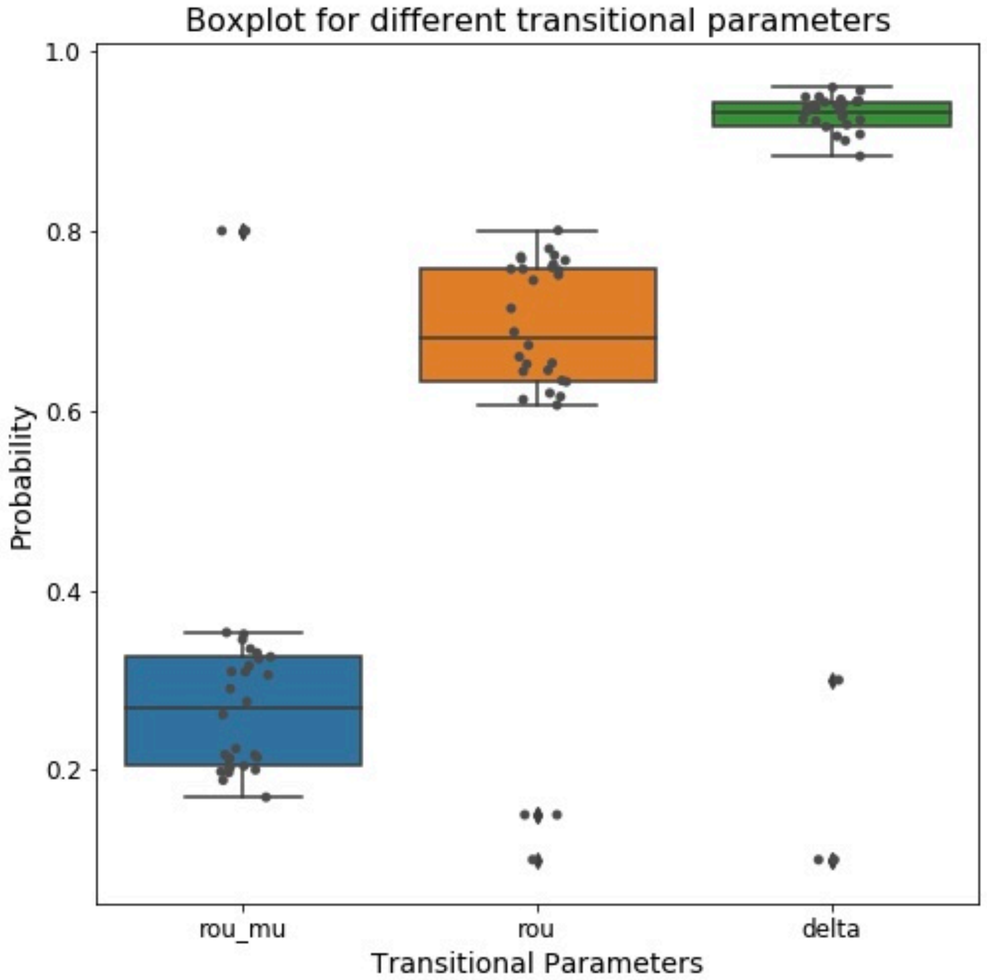
Start codon: AUG

Simulation Results

Simulate 30 datasets, each contains 100 RNA sequences
Start codon: AUG
Stop codon: UAA, UGA, UAG

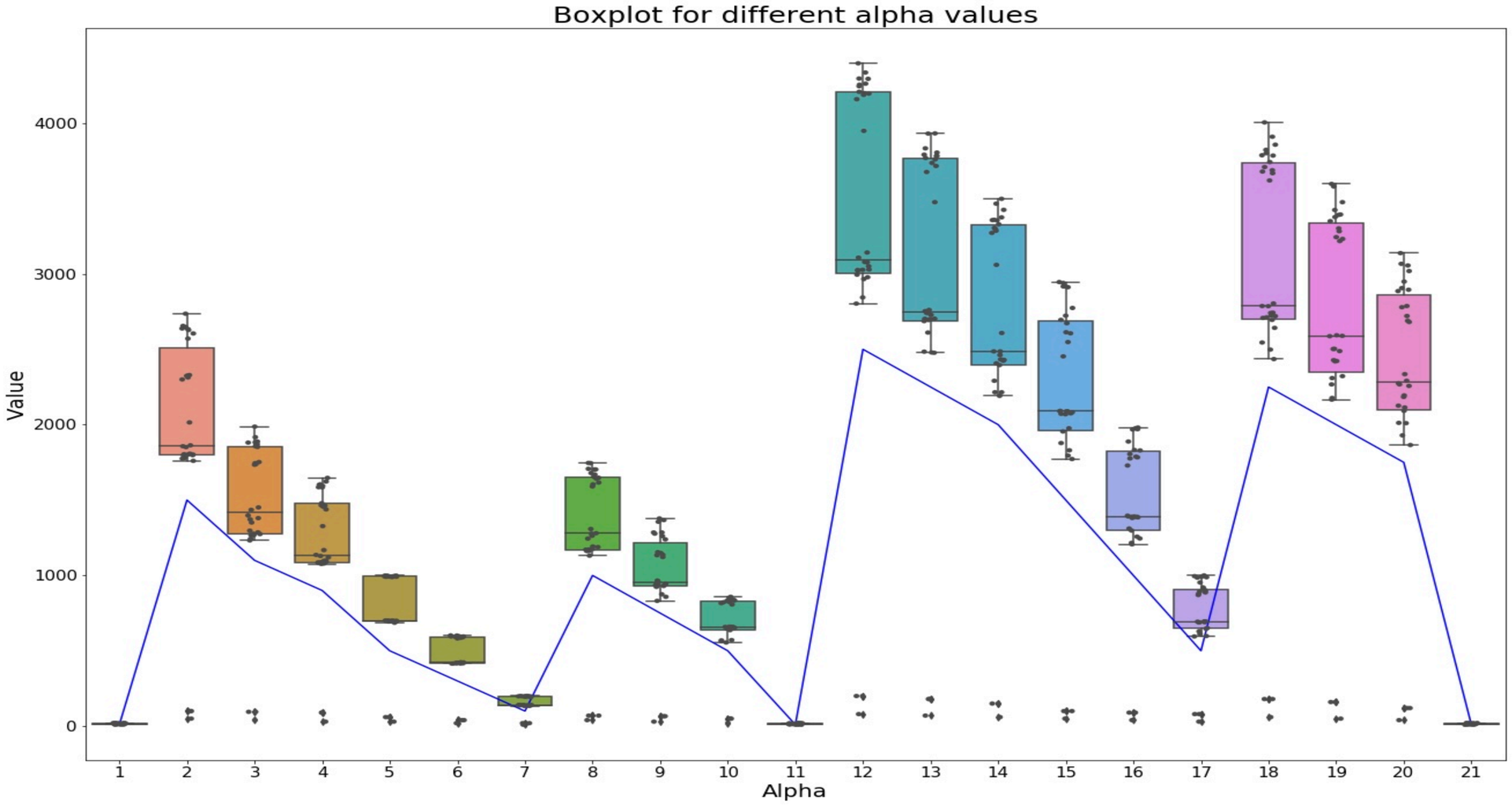
Simulation Results

Boxplot for transitional parameters after EM algorithm
Specific start codon: AUG



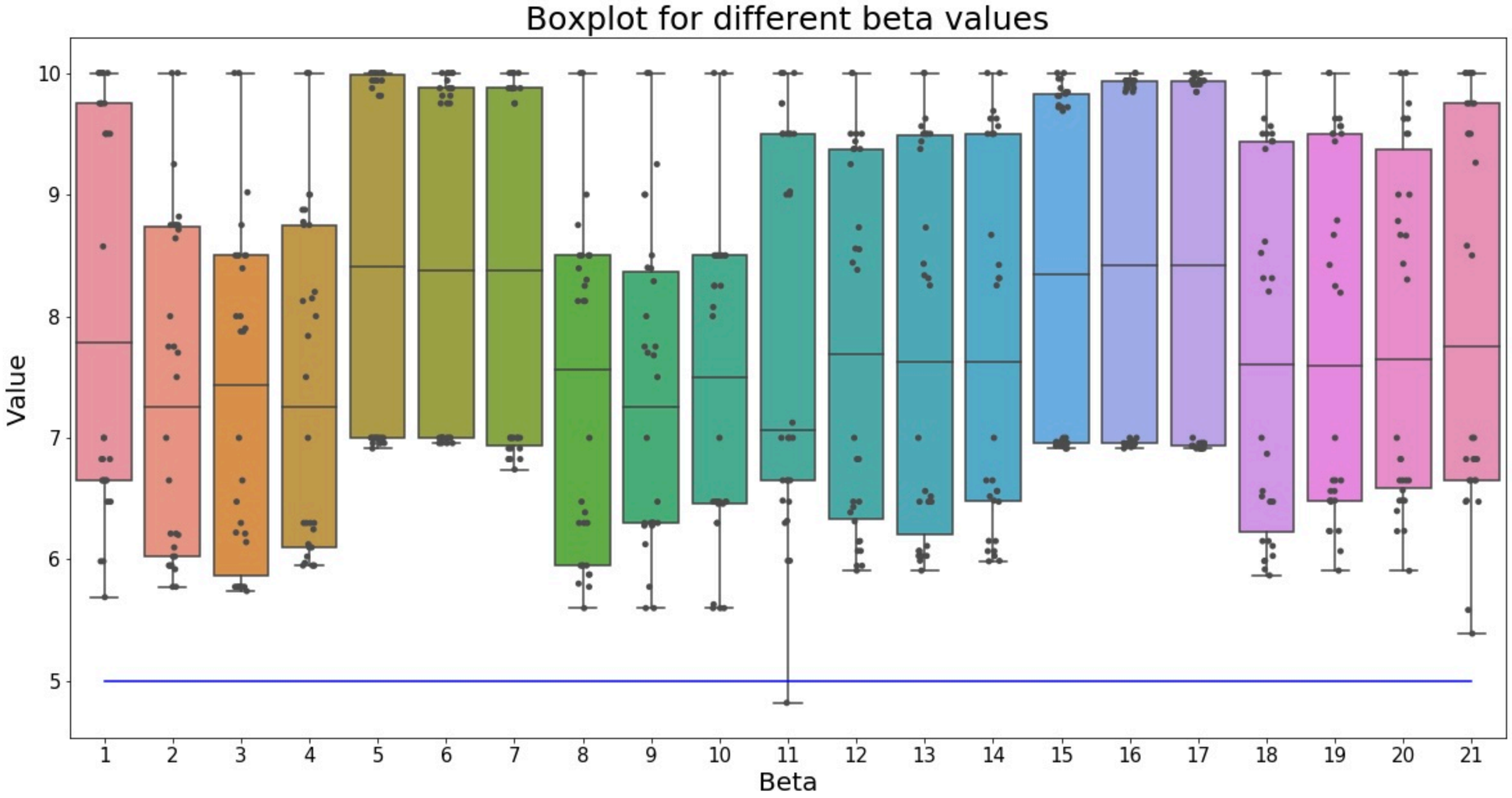
Simulation Results

Boxplot for emission parameter α_z after EM algorithm



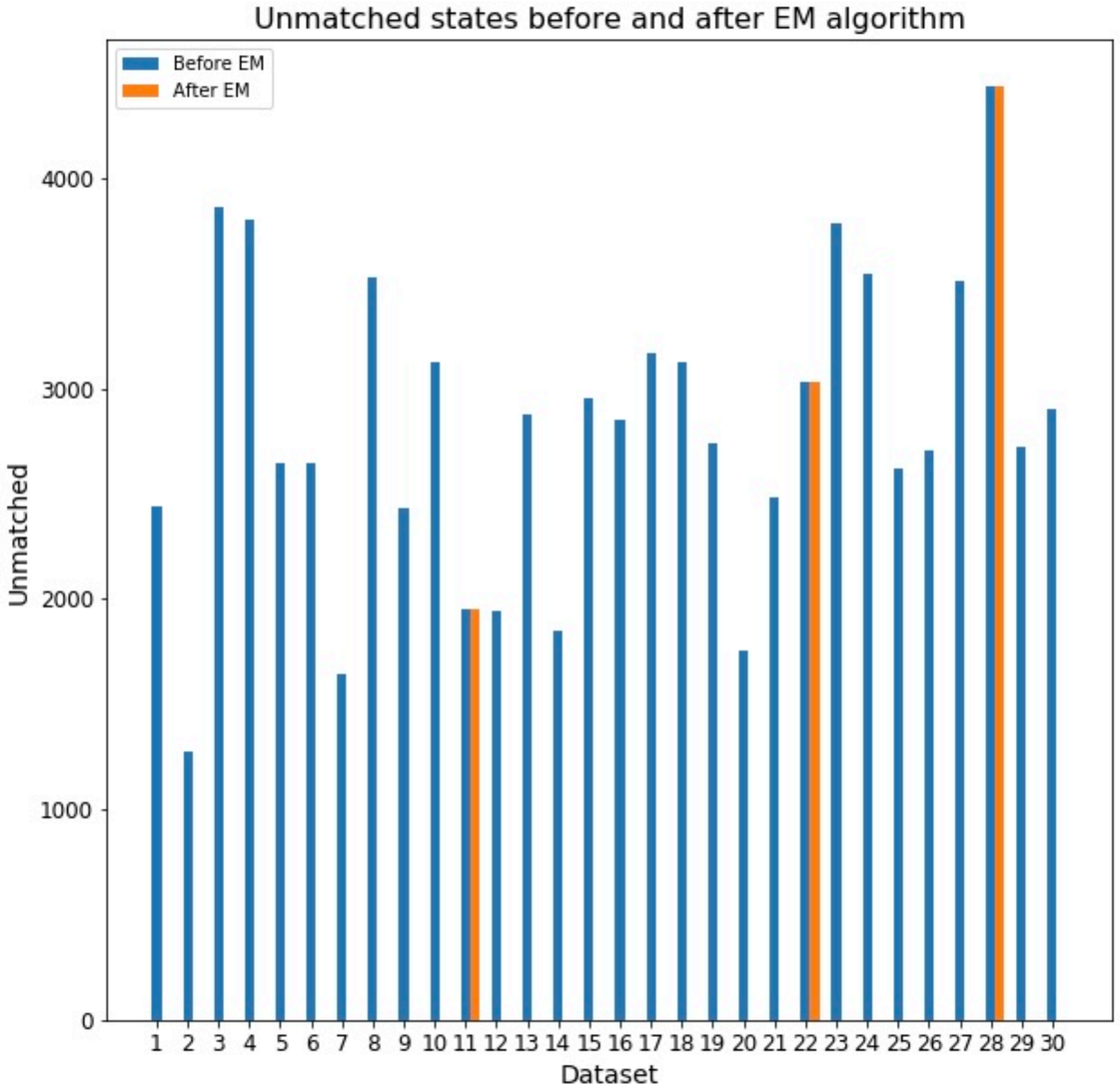
Simulation Results

Boxplot for emission parameter $\boldsymbol{\beta}_Z$ after EM algorithm



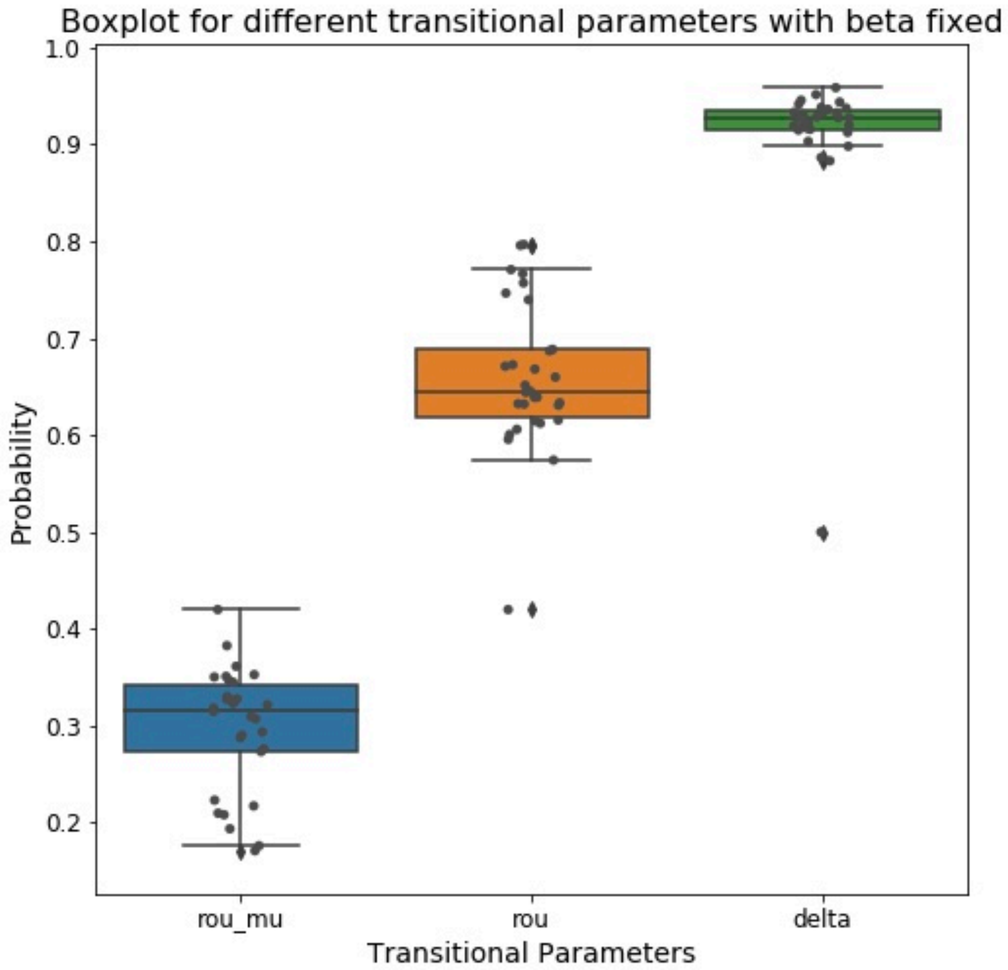
Simulation Results

Number of unmatched states before and after EM algorithm



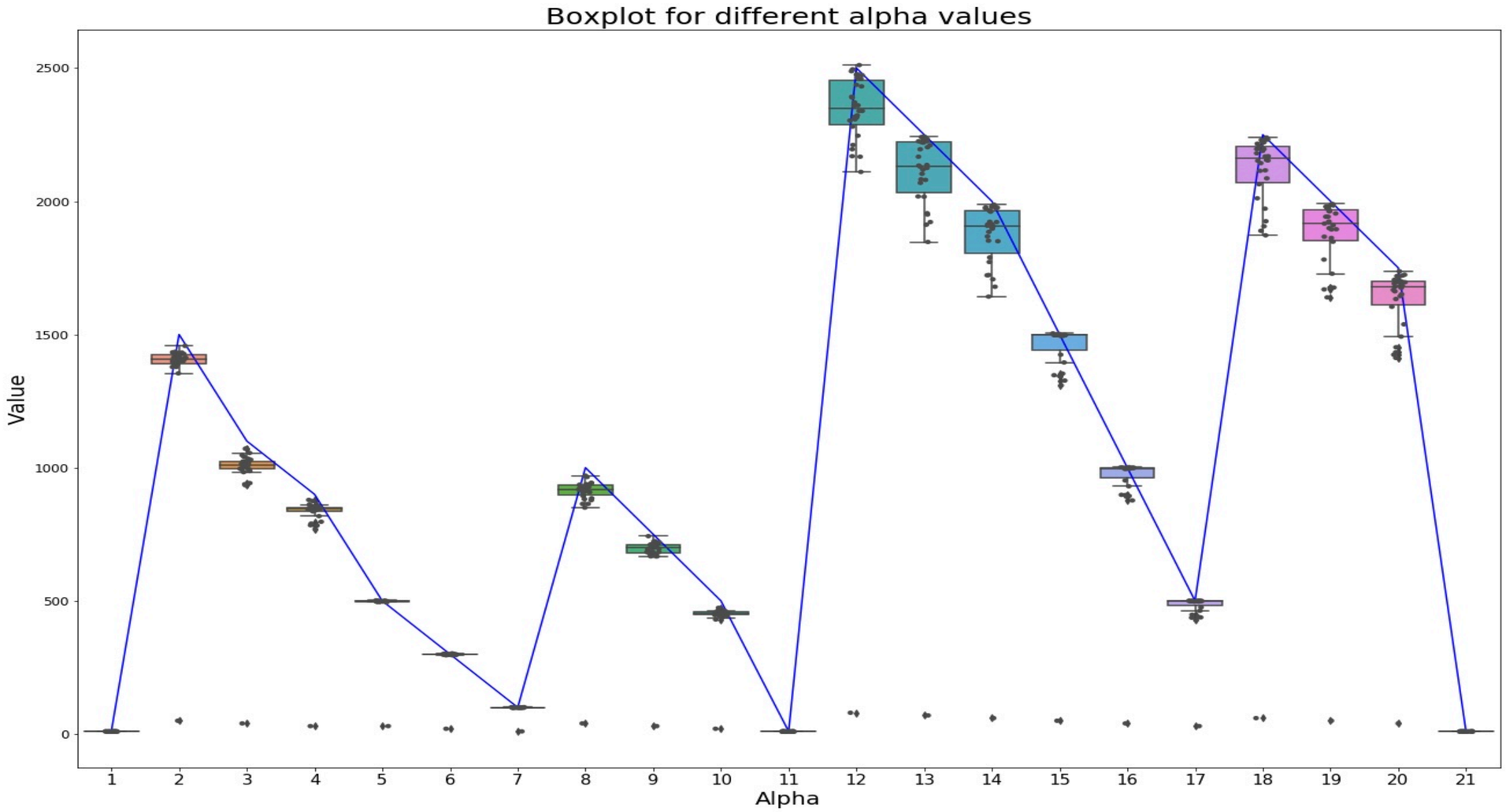
Simulation Results - $\boldsymbol{\beta}_z$ Fixed

Boxplot for transitional parameters after EM algorithm
Specific start codon: AUG



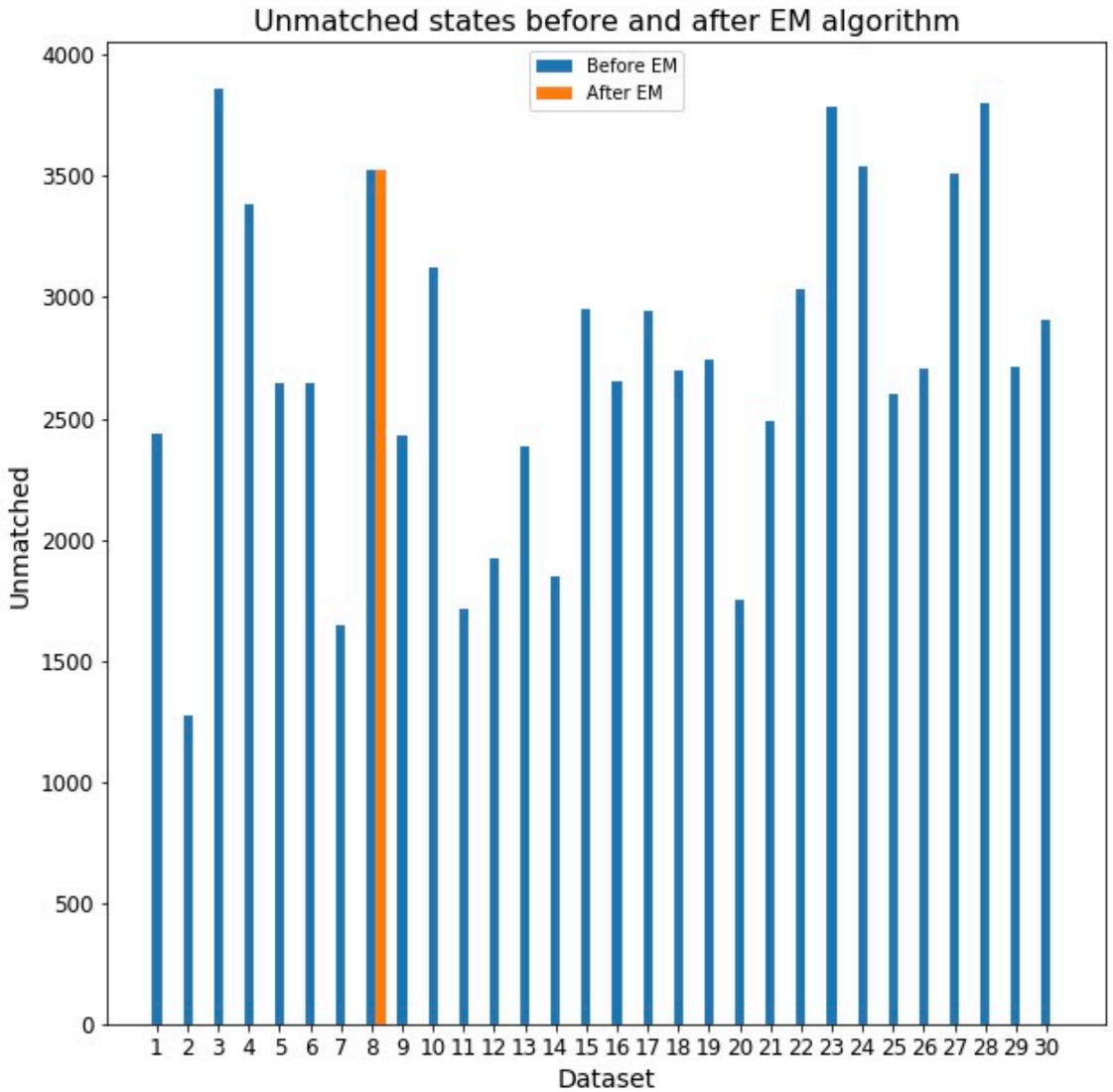
Simulation Results - $\boldsymbol{\beta}_Z$ Fixed

Boxplot for emission parameter $\boldsymbol{\alpha}_Z$ after EM algorithm



Simulation Results - $\boldsymbol{\beta}_Z$ Fixed

Number of unmatched states before and after EM algorithm



Current Work

1. Identified translated uORFs and main coding regions with Hidden Markov Model and EM algorithm

Current Work

1. Identified translated uORFs and main coding regions with Hidden Markov Model and EM algorithm
2. Implemented this model in Python

Current Work

1. Identified translated uORFs and main coding regions with Hidden Markov Model and EM algorithm
2. Implemented this model in Python
3. Tested the model with simulation data

Future Work

1. Improve parameter estimation with Automatic Differentiation

Future Work

1. Improve parameter estimation with Automatic Differentiation
2. Apply this model on real biological data

Thanks

Jiayuan Zhu
Supervisor: Heejung Shim
Acknowledgement: Jiangrong Ouyang's previous work

4 June 2021