# Simulation of mtDNA Transcription with Gillespie Algorithm

Alex Yang

August 2024

## 1 Gillespie Algorithm

### 1.1 Stochastic Simulation Algorithm (SSA)

Gillespie algorithm originated as a method to simulate multiple chemical reactions. Using ODE to simulate the reactions could be computationally expansive. Instead, SSA divides the time line into multiple time segments. Each segment is short enough such that it is safe to assume that only one reaction take place in it. Therefore, one only needs to care about the time till the next reaction (i.e. length of the time segment) and the next reaction index (i.e. which reaction to take place), and update the amount of the chemicals $X(t)$ once per time segment.

To obtain these two key features, we introduce the propensity function $a_j(X(t))$ associated with reaction $j$, which satisfies that the probability for the reaction to take place in time interval $[t, t + dt)$ is given by $a_j(X(t))dt$. The construction of propensity function for certain reaction depends on the type of the reaction (i.e. how many components attend in the reaction) and the amount of each components available for reacting.

For reaction $j$, the time till next reaction is exponentially distributed with density function $a_j(x)e^{-a_j(x)}$. Since each reaction happens independently, the time till any reaction happen is also exponentially distributed with the density function to be $a_{sum}(x)e^{-a_{sum}(x)}$. Meanwhile, since the probability of picking a reaction is proportional to its propensity function, the next reaction index can be drawn from a discrete random variable with density function $a_j(x)/a_{sum}(x)$.

### 1.2 Tau-Leaping

SSA is exact but still computationally expansive. In the case that some reactions happen rapidly and/or some chemicals comes in large amount, the 'time till next reaction' in SSA could be typically small, which means lots of random variables need to be drawn.

Instead, we now consider suitably longer time segments $\tau$ to allow multiple reactions taking place. We will try to approximate the number of each reaction based on the data at the beginning of the time segment, i.e. we freeze the propensity function within the segment. This gives the tau-leaping method

$$X(t+\tau) = X(t) + \sum_{M}^{j=1} \nu_j P_j(a_j(X(t)), \tau), \tag{1}$$

where $P_j(a_j(X(t)), \tau)$ is the number of reactions $j$ happening, and $\nu_j$ represent how the reaction works. This method is valid only if the length of the segment is sufficiently small such that the amount of the chemicals is not significantly changed.

Using a proper time length, we may further assume that the amount of the components are constant within the segment, in which case the number of each reaction will follow Poisson distribution $Poi(a_j(X(t)))$.

# 2 Simulation on mtDNA Transcription

## 2.1 Background

The project aims to use Gillespie algorithm to simulate the Mitochondrial DNA Transcription. There are 37 mt-genes in human mitochondria DNA, and we will focus on the transcription of the 13 protein coding genes and the 2 rRNAs.

Transcription of the genes can be initiated either on the light strand or on the heavy strand. On the light strand, it will only replicate one protein-coding gene (ND6). On the heavy strand, it starts from replicating the two rRNAs, and has a chance to terminate right after that; otherwise, it will continue to replicate all the remaining 12 protein-coding genes.

Meanwhile, the gene transcripts have a chance to degrade at time. The quantity of transcripts of each gene is subject to exponential decay, where the explicit decay coefficient is unknown. In the simulation, we will assume decay coefficients for the transcripts, and use the algorithm to solve for the stable status.

## 2.2 Assumptions

We make following assumptions to apply numerical simulation on the transcription process in a typical sized cell:

- We assume that the transcription has constant initiation rates within an mtDNA, which is $p_h$ on the heavy strand and $p_l$ on the light strand.
- A proportion of transcription over the heavy strand may terminate after transcribing the two rRNAs. We set this proportion to be 10%.
- During the simulation, we further assume that the initiation rates are $p_h = p_l = 20/\text{min}$, and there are 100 mtDNAs within a typical-sized cell.

- A decay constant is assumed for each of the rRNA / protein coding gene to control the rate of degradation. The constants are chosen based on the result of the following single cell sequencing experiment.

## 2.3 Decay Constants

Single cell sequencing was applied on samples from two donors (including two cell types: CD4+ and CD8+ T cells). The datasets includes counts for all the 37 mt-gene transcripts and the non-mt gene transcripts, where the later one can indicate the size of the observed sample. To normalize the data, we divide the data of each sample by its non-MT count. Then we average the data across the samples of a cell type to obtain the typical amount of gene transcripts per non-mt gene transcript.

At current stage we wish to obtain simulation results where each gene transcript takes the same proportion with the above dataset. Indeed, we can approximate the simulation results by ODE, though they are not stochastic. The quantity $N_i$ of transcripts of the $i$th gene can be described as

$$\frac{\mathrm{d}N_i}{\mathrm{d}t} = -\lambda_i N_i + P_i, \tag{2}$$

where $\lambda_i$ is the decay constant to be determined. The production rate $P_i$, by assumption, equals to 2000/min for ND6, 2000/min for the two rRNA, and 1800/min for the twelve protein coding genes on the heavy strand.

The quantity $N_i$ is stable when $N_i = P_i/\lambda_i$. Hence if we fix the ratio between $N_i$ at the stable status in compliance with the experiment data, the decay constants $\lambda_i$ can be determined up to scaling.

| Name | Decay Constant | Half-life |
|---|---|---|
| MT-RNR1 | 0.003227 | 3.579696 |
| MT-RNR2 | 0.000870 | 13.272049 |
| MT-ND1 | 0.009268 | 1.246539 |
| MT-ND2 | 0.015645 | 0.738430 |
| MT-CO1 | 0.003640 | 3.173496 |
| MT-CO2 | 0.002137 | 5.406789 |
| MT-ATP8 | 0.003767 | 3.067105 |
| MT-ATP6 | 0.034096 | 0.338823 |
| MT-CO3 | 0.002719 | 4.248379 |
| MT-ND3 | 0.012971 | 0.890630 |
| MT-ND4L | 0.005564 | 2.076371 |
| MT-ND4 | 0.021766 | 0.530763 |
| MT-ND5 | 0.014269 | 0.809600 |
| MT-ND6 | 0.077088 | 0.149860 |
| MT-CYB | 0.003476 | 3.323214 |

Table 1: Assumed decay constants (in $\text{min}^{-1}$) and half-life (in hour) for D1CD4

In addition, the half-life of the transcripts are expected to be around 3.8 days to 7.5 days for the rRNAs and around 1-2 hours for the mRNAs. In practice, we scale the decay constants such that the average resulting half life of the 13 protein-coding transcripts is 2 hours. The assumed decay constants processed from the samples of CD4+ from Donor 1 are shown in the Table 1 for instance.

## 2.4 Results

We applied the algorithm and simulated the transcription and degradation process. For each set of decay constant, we will iterate the algorithm starting from zero transcripts for 1000 times, and take the average as the results. The step size is set to be one minute per step, and the simulation for each set of constants takes around 30000 steps to become about stable and terminate.

A set of decay constants is chosen for each type of cell from each donor. For each row of figures, the blue bars show the result of the simulation, and the red bars show the distribution of corresponding samples as comparison. The decay constants in the last row are obtained from the average across the four cell types.

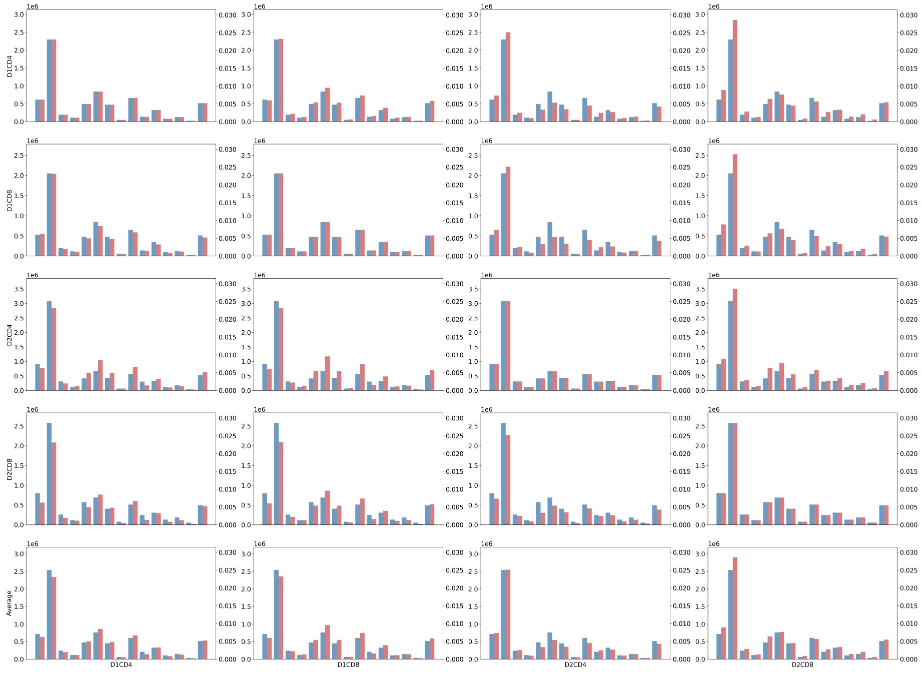

Figure 1: Comparison between distributions of simulation results (in blue) and the processed sc-seq data (in red)

We also obtained the coefficient of variation of the results, shown in the figures below in blue. As comparison, the red bars show the coefficient of variation of the raw data obtained in the single cell sequencing.

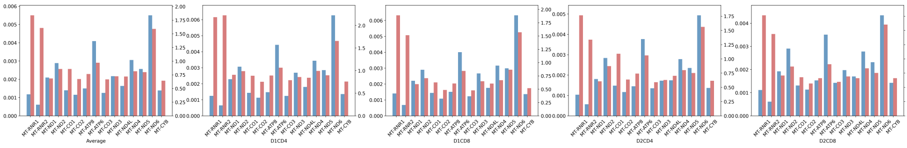

Figure 2: Comparison between the coefficient of variation of the results and the raw data

We further compare the results above with data obtained from smart single cell sequencing. Due to difference in experiment process, the simulation result does not match properly with the later one.

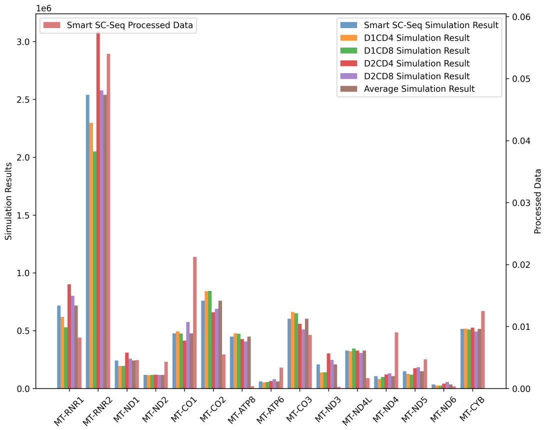

Figure 3: Comparison of simulation results from all sets of constants and the processed data from smart sc-seq

Finally, all the decay constants assumed are listed in the table below:

| Name | D1CD4 | D1CD8 | D2CD4 | D2CD8 | Average | Smart SC-Seq |
|---|---|---|---|---|---|---|
| MT-RNR1 | 0.003227 | 0.003777 | 0.002218 | 0.002498 | 0.002855 | 0.004611 |
| MT-RNR2 | 0.000870 | 0.000976 | 0.000648 | 0.000776 | 0.000812 | 0.000701 |
| MT-ND1 | 0.009268 | 0.009235 | 0.005809 | 0.006959 | 0.007653 | 0.007443 |
| MT-ND2 | 0.015645 | 0.015399 | 0.015119 | 0.015403 | 0.015403 | 0.007858 |
| MT-CO1 | 0.003640 | 0.003781 | 0.004340 | 0.003133 | 0.003635 | 0.001602 |
| MT-CO2 | 0.002137 | 0.002133 | 0.002726 | 0.002611 | 0.002355 | 0.006175 |
| MT-ATP8 | 0.003767 | 0.003796 | 0.004209 | 0.004426 | 0.004024 | 0.097574 |
| MT-ATP6 | 0.034096 | 0.031806 | 0.027466 | 0.022443 | 0.028172 | 0.010064 |
| MT-CO3 | 0.002719 | 0.002768 | 0.003212 | 0.003523 | 0.003014 | 0.003942 |
| MT-ND3 | 0.012971 | 0.012846 | 0.005919 | 0.007274 | 0.008912 | 0.124246 |
| MT-ND4L | 0.005564 | 0.005204 | 0.005474 | 0.005870 | 0.005518 | 0.020404 |
| MT-ND4 | 0.021766 | 0.018099 | 0.014864 | 0.013712 | 0.016615 | 0.003752 |
| MT-ND5 | 0.014269 | 0.015036 | 0.010210 | 0.009728 | 0.011961 | 0.007211 |
| MT-ND6 | 0.077088 | 0.081432 | 0.047655 | 0.035394 | 0.053884 | 0.103111 |
| MT-CYB | 0.003476 | 0.003524 | 0.003425 | 0.003654 | 0.003526 | 0.002720 |

Table 2: All assumed decay constants (in $\text{min}^{-1}$)