Transcription

Model implementation. The *E. coli* model assumes RNA polymerase exists in two states: free and actively transcribing. Every time step, free RNA polymerase transitions to the actively transcribing state to maintain an experimentally-observed active fraction of RNA polymerase. This is a simplification compared to *M. genitalium* model, which modeled RNA polymerase as existing in 4 states: free, non-specifically bound on a chromosome, bound to a promoter, and actively transcribing a gene. The *E. coli* model does not yet include sigma, elongation or termination factors. The *E. coli* model also currently treats each gene as its own transcription unit. Transcription occurs through the action of two processes in the model: `TranscriptInitiation` (Algorithm 1) and `TrancriptElongation` (Algorithm 2).

Initiation. `TranscriptInitiation` models the binding of RNA polymerase to each gene. The number of initiation events per gene is determined in a probabilistic manner and dependent on the number of free RNA polymerases and each gene's synthesis probability. The number of RNA polymerases to activate in each time step is determined such that the average fraction of RNA polymerases that are active throughout the simulation matches measured fractions, which are dependent on the cellular growth rate. This is done by assuming a steady state concentration of active RNA polymerases (and therefore a constant active fraction):

$$\frac{dR_{act}}{dt} = p_{act} \cdot R_{free} - r \cdot R_{act} = 0 \tag{1}$$

$$p_{act} = \frac{r \cdot R_{act}}{R_{free}} \tag{2}$$

where R_{act} is the concentration of active RNA polymerases, R_{free} is the concentration of free RNA polymerases, p_{act} is the activation probability and r is the expected termination

rate for active RNA polymerases. Using the definition of the active fraction, $f_{act} = \frac{R_{act}}{R_{act}+R_{free}}$, p_{act} can be defined in terms of the desired active fraction:

$$p_{act} = \frac{r \cdot f_{act}}{1 - f_{act}} \tag{3}$$

This activation probability is then used to determine how many free RNA polymerases will initiate. These newly initiated RNA polymerases are distributed to individual genes based on the synthesis probability for each gene, which is determined based on another steady state assumption for each mRNA concentration:

$$\frac{dm_i}{dt} = v_{synth,i} - m_i \cdot \left(\frac{\ln 2}{\tau} + \frac{\ln 2}{t_{\frac{1}{2},i}}\right) = 0 \tag{4}$$

$$v_{synth,i} = m_i \cdot \left(\frac{\ln 2}{\tau} + \frac{\ln 2}{t_{\frac{1}{2},i}}\right) \tag{5}$$

where $v_{synth,i}$ is the synthesis rate of each mRNA, m_i is the concentration of each mRNA, τ is the doubling time and $t_{\frac{1}{2},i}$ is the half life for each mRNA (see Section **??**). Using RNA expression data for m_i, the rate of synthesis for each gene can be determined. Synthesis rates are then normalized as below to determine a synthesis probability for each gene:

$$p_{synth,i} = \frac{v_{synth,i}}{\sum_j v_{synth,j}} \tag{6}$$

where $p_{synth,i}$ is the synthesis probability for each gene. Gene synthesis probabilities are further dependent on transcription factor binding and regulation as discussed in the next section (Section **??**).

Elongation. `TranscriptElongation` models nucleotide polymerization into RNA molecules by RNA polymerases. Polymerization occurs across all polymerases simultaneously and resources are allocated to maximize the progress of all polymerases up to the limit of the expected polymerase elongation rate and available nucleotides. The termination of RNA elongation occurs once a RNA polymerase has reached the end of the

annotated gene.

Model assumptions. The *E. coli* genome contains seven copies of the rRNA operon, and all seven copies contribute to the transcription of ribosomal RNAs. The sequences of the rRNA genes in these operons are known to be slightly different from one another, but it is unclear whether these small differences in sequence lead to significant functional differences between these molecules [2]. In our model, we make the assumption that all seven rRNA operons produce rRNAs that have sequences identical to rRNAs from the *rrnA* operon, such that there is only a single representation for each type of rRNA molecule (23S, 16S, 5S) inside the model. This significantly simplifies the modeling of the complexation reactions that produce the ribosomal subunits, as we do not need to consider all combinations of rRNAs that may be complexed together to form distinct ribosomal subunits.

Algorithm 1: RNA polymerase initiation on DNA

Input : f_{act} fraction of RNA polymerases that are active
Input : r expected termination rate for active RNA polymerases
Input : $p_{\text{synth,i}}$ RNA synthesis probability for each gene where $i = 1$ **to** n_{gene}
Input : $c_{RNAP,f}$ count of free RNA polymerase
Input : `multinomial()` function that draws samples from a multinomial distribution

1. Calculate probability (p_{act}) of a free RNA polymerase binding to a gene.

$$p_{act} = \frac{r \cdot f_{act}}{1 - f_{act}}$$

2. Calculate the number of RNA polymerases that will bind and activate ($c_{RNAP,b}$).

$$c_{RNAP,b} = p_{act} \cdot c_{RNAP,f}$$

3 Sample multinomial distribution $c_{RNAP,b}$ times weighted by $p_{synth,i}$ to determine which genes receive a RNA polymerase and initiate ($n_{init,i}$).

$$n_{init,i} = \texttt{multinomial}(c_{RNAP,b}, p_{synth,i})$$

4 Assign $n_{init,i}$ RNA polymerases to gene i. Decrement free RNA polymerase counts.

Result: RNA polymerases bind to genes based on the number of free RNA polymerases and the synthesis probability for each gene.

Algorithm 2: mRNA elongation and termination

Input : e expected RNA polymerase elongation rate in given environment

Input : L_i length of each gene $i = 1$ **to** n_{gene} for each coding gene.

Input : p_j gene position of RNA polymerase $j = 1$ **to** n_{RNAP}

Input : $c_{nt,k}$ counts of nucleotides $k = 1$ **to** 4 for each nucleotide type (A, C, G, U)

Input : Δt length of current time step

`/* Elongate RNA transcripts up to limits of sequence or nucleotides */`

for *each RNA polymerase j on gene i* **do**

1. Based on RNA polymerase position p_j on a gene i and maximal elongation rate e determine "stop condition" (s_j) for RNA polymerase j assuming no nucleotide limitation.

$$s_j = \texttt{min}(p_j + e \cdot \Delta t,\ L_i)$$

Stop condition is either maximal elongation rate scaled by the time step or the full length of sequence (i.e. the RNA polymerase will terminate in this time step).

2. Derive sequence between RNA polymerase position (p_j) and stop condition (s_j).

3. Based on derived sequence calculate the number of nucleotides required to polymerize sequence $c^{req}_{nt,k}$.

4. Elongate up to limits:

if `all`$(c^{req}_{nt,k} < c_{nt,k})$ **then**

Update the position of each polymerase to stop position

$$p_j = s_j$$

else

4a. Attempt to elongate all RNA fragments.

4b. Update position of each polymerase to maximal position given the limitation of $c_{nt,k}$.

5. Update counts of $c_{nt,k}$ to reflect polymerization usage.

`/* Terminate RNA polymerases that have reached the end of their gene */`

for *each RNA polymerase j on gene i* **do**

if $p_j == L_i$ **then**

1. Increment count of RNA that corresponds to elongating RNA transcript that has terminated.

2. Increment free RNA polymerase counts.

Result: Each RNA transcript is elongated up to the limit of available gene sequence, expected elongation rate, or nucleotide limitation. RNA polymerases that reach the end of their genes are terminated and released.

Associated data

Parameter	Symbol	Units	Value	Reference
Active fraction of RNAP	f_{act}	-	[0.17, 0.30]	[1]
RNA synthesis probability	p_{synth}	-	[0, 0.015]	See GitHub
RNAP elongation rate	e	nt/s	[39, 55]	[1]

Table 1: Table of parameters for Transcript Initiation and Elongation processes.

Associated files

`wcEcoli` Path	File	Type
`wcEcoli/models/ecoli/processes`	`transcript_initiation.py`	process
`wcEcoli/models/ecoli/processes`	`transcript_elongation.py`	process
`wcEcoli/reconstruction/ecoli/dataclasses/process`	`transcription.py`	data
`wcEcoli/reconstruction/ecoli/flat`	`rnas.tsv`	raw data
`wcEcoli/reconstruction/ecoli/flat`	`growthRateDependentParameters.tsv`	raw data

Table 2: Table of files for transcription.

References

[1] Hans Bremer and Patrick P Dennis. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3(1), 2008.

[2] Michihisa Maeda, Tomohiro Shimada, and Akira Ishihama. Strength and regulation of seven rrna promoters in escherichia coli. *PloS one*, 10(12):e0144697, 2015.