

## History

- Initial model by mikeg on 6/10/15.
- NSE model added by mikeg on 7/21/15.
- Fixed error for  $\Pr(\text{Elongation at position } j)$  in Equation (13) by adding  $\lambda_j v_{g,i}$  term. Additionally, applied fix to Equation (16) by adding missing term to product term. Added notes about expecting  $\alpha_c > 0$ .
- Compiled on Friday 17<sup>th</sup> August, 2018 at 17:23

## Goal

Create sensible model for interpreting RPF data with a special interest in working with data used in Pop et al. (2014).

## Pausing Time Model Definition

### Calculating the Likelihood of a sample

We are interested in calculating the probability of observing a ribosome footprint (RFP) experimentally. We assume there is a pool of RFP generated from the transcriptome, that the mRNAs in this pool are at close to steady state in terms of ribosome initiation and completion of translating a transcript.

Beginning by considering a single mRNA molecule transcribed from gene  $g$ , the probability a ribosome is at position  $i$  of this mRNA molecule,  $p_{g,i}$ , is simply,

$$p_{g,i} \propto \kappa_g w_{g,i} \tag{1}$$

where  $\kappa_g$  is a rate constant scales the average rate at which ribosomes intercept and initiate translation of an mRNA molecule from gene  $g$ , mRNA $_g$ , under the experimental conditions used. Formally, the initiation rate is determined by  $\kappa_g \times r$ , where  $r$  is the density of ribosomes in the cell. However, we assume all mRNAs are equally accessible to ribosomes so, as a result,  $r$  will cancel out in the following equation and, as a result, we ignore it throughout. Additionally,  $w_{g,i}$  is the average waiting, pausing, or dwell time of a ribosome at position  $i$  of mRNA from gene  $g$ . Derivation of equation 1 is straight forward, but can also be found Gilchrist and Wagner (2006) equation (20) with the nonsense error rate = 0 and the ribosome recycling probability

$< 1$ . We can link it to [Pop et al. \(2014\)](#) work by noting the ribosome flux  $J_g$  on an individual mRNA $_g$  is  $J_g = r\kappa_g$ .

Biologically speaking,  $w_{g,i}$  values for the same codon are not independent. The values of  $w_{g,i}$  for the relevant codon likely vary within a gene as a function of mRNA structure and other factors. To capture this variation, we will assume that for when the codon at position  $i$  is of type  $c$ ,  $w_{g,i} \sim \text{Gamma}(\alpha_c, \lambda_c)$ , where  $\alpha_c$  is the shape parameter,  $\lambda_c$  is the rate parameter, and  $E[w_{g,i}] = \alpha_c/\lambda_c$ . Gene specific effects could also be incorporated since mRNA structures which interfere with efficient translation likely declines with expression level. As a result of this assumption, we can use a Negative-Binomial (NB) distribution model to analyze the RFP data. Regarding values of  $\alpha_c$ , no variation in waiting times between instances of the same codon would correspond to an exponential model or  $\alpha_c = 1$ . Given that we expect heterogeneity between sites, we expect  $\alpha_c > 1$ .

Assuming independence in sampling, the total probability of a randomly selected footprint is from position  $i$ ,  $P_{g,i}$ , is

$$P_{g,i} = p_{g,i}m_g/Z \quad (2)$$

where  $m_g$  is the density of mRNA $_g$  in the cell and  $Z$  is a partition function that ensures our sampling probabilities across the transcriptome sums to 1 and is defined as,

$$Z = \sum_g m_g \left( \sum_i p_{g,i} \right). \quad (3)$$

Equations (2) and (3) indicate that our choices of time and volume units for  $\kappa_g$  and  $m_g$  are irrelevant and we can only estimate their values relative to one another.

Letting  $Y = \sum_{g,i} Y_{g,i}$  be the footprint sample size, where  $Y \gg 1$  and  $P_{g,i} \ll 1$ , then we can approximate the probability of observing  $Y_{g,i}$  samples of a footprint in mRNA $_g$  at position  $i$  using a Poisson distribution with a sampling rate of  $YP_{g,i}$ . Note that deep sequencing may violate the sampling with replacement assumption of the Poisson distribution. Given our assumption about the distribution of  $p_{g,i}$ ,  $Y_{g,i} \sim \text{NB}(x = \alpha_c, p = \kappa_g m_g / (\lambda'_c + \kappa_g m_g))$  where  $\lambda'_c = \lambda_c Z / Y$ . Explicitly,

$$\begin{aligned} \Pr(Y_{g,i} | \alpha_c, \lambda'_c, m_g, \kappa_g) &= \frac{\Gamma(\alpha_c + Y_{g,i})}{\Gamma(\alpha_c) Y_{g,i}!} \left( \frac{m_g \kappa_g}{\lambda'_c + m_g \kappa_g} \right)^{Y_{g,i}} \left( \frac{\lambda'_c}{\lambda'_c + m_g \kappa_g} \right)^{\alpha_c} \\ &= \frac{\Gamma(\alpha_c + Y_{g,i})}{\Gamma(\alpha_c) Y_{g,i}!} \left( \frac{m_g \kappa_g}{\lambda'_c + m_g \kappa_g} \right)^{Y_{g,i}} \left( 1 - \frac{m_g \kappa_g}{\lambda'_c + m_g \kappa_g} \right)^{\alpha_c} \end{aligned} \quad (4)$$

Note that  $m_g$  and  $\kappa_g$  are gene  $g$  and environment specific terms which can be equated to the equilibrium

protein synthesis initiation rate  $\psi_g$  for gene  $g$  under the experimental conditions, i.e.  $\psi_g = \kappa_g m_g$ . Further, because in this model nonsense errors are ignored, all ribosomes that initiate translation also complete translation. Thus, the initiation rate  $\psi_g$  is also equal to the synthesis rate  $\phi_g$ . The composite parameter  $\lambda'_c$  consists of the codon specific scale term  $\lambda_c$  and the ratio of  $Z$  to  $Y$ , two genome wide parameters. The term  $Y/Z$  can be interpreted as the sampling efficiency, i.e. what proportion of the RFP state space is sampled. If  $Y/Z \ll 1$ , then sampling is sparse.

Given the properties of the NB (Forbes et al., 2011, p. 141) and the fact that most codons appear within a gene's ORF multiple times, the likelihood of the parameters given the total number of RFP observed derived from codons of type  $c$  in gene  $g$ ,  $Y_g^c = \sum_{i \in c} Y_{g,i}$ , is,

$$L(\psi_g, \alpha_c, \lambda'_c | Y_g^c, n_g^c) = \frac{\Gamma(n_g^c \alpha_c + Y_g^c)}{\Gamma(n_g^c \alpha_c)} \left( \frac{\psi_g}{\lambda'_c + \psi_g} \right)^{Y_g^c} \left( 1 - \frac{\psi_g}{\lambda'_c + \psi_g} \right)^{n_g^c \alpha_c} \quad (5)$$

$$= \frac{\Gamma(\alpha'_{c,g} + Y_g^c)}{\Gamma(\alpha'_{c,g})} \left( \frac{\psi_g}{\lambda'_c + \psi_g} \right)^{Y_g^c} \left( 1 - \frac{\psi_g}{\lambda'_c + \psi_g} \right)^{\alpha'_{c,g}} \quad (6)$$

where  $n_g^c$  is the number of times codon  $c$  is found in the ORF of gene  $g$  and  $\alpha'_{c,g} = n_g^c \times \alpha_c$ . Note we dropped the  $Y_g^c!$  term because that will be the same for all parameter values. The total Likelihood of the data is

$$L(\vec{\psi}_g, \vec{\alpha}_c, \vec{\lambda}'_c | \vec{Y}_g^c, \vec{n}_g^c) = \prod_{g \in \mathbb{G}} \prod_{c \in \mathbb{C}} L(\psi_g, \alpha_c, \lambda'_c | Y_g^c, n_g^c) \quad (7)$$

Note that using RFP data alone,  $\kappa_g \times m_g = \psi_g$  and  $\lambda'_c = \lambda_c Z/Y$  are only identifiable as joint parameters. (Although it seems like you should be able to calculate  $Z$  post-hoc from the state of the chain and estimates of  $m_g$  are available from other sources.) Most standard libraries require that the  $x$  parameter in a NB be discrete, which in this case it is not. Thus to simulate  $Y_g$  values based on equations (4) or (6), first pull  $W$  from  $\text{Gamma}(n_g^c \alpha_c, \lambda'_c)$ <sup>1</sup>, then pull  $Y$  from  $\text{Poisson}(W \psi_g)$ .

Pop et al. (2014) provide RNA-Seq based counts of mRNA abundances,  $M_g$ , in addition to RFP counts. If we assume that  $M_g \sim \text{Poisson}(Y m_g)$ , we can easily combine both the RFP and the mRNA counts together and estimate  $\kappa_g$  and  $m_g$  separately. Alternatively, we could estimate the composite  $\kappa_g m_g$  parameter using the RFP data and then analyze the those results using the  $M_g$  data.

An additional fact worth noting is that we are treating  $Z$  as a constant for much of our calculations when, in fact, it is a random variate. We should discuss this with Russ to ensure there are no issues.

---

<sup>1</sup>recall that  $\lambda'_c$  is the 'rate' parameter

## Pausing Time with Nonsense Error Model Definition

The flux equation, Equation (1), no longer holds when nonsense errors (NSE) are possible. Instead, following Gilchrist and Wagner (2006), we have the conditional probability,

$$p_{g,i}|w_{g,i} \propto \kappa_g \sigma(i-1) \frac{w_{g,i} v_{g,i}}{w_{g,i} + v_{g,i}} \quad (8)$$

Where, as before,  $\kappa_g$  is the translation initiation rate constant per mRNA<sub>*g*</sub>,  $w_{g,i}$  is the waiting time to elongate codon *i*, and  $w_{g,i}$  is 1 over the codon elongation rate ( $1/c_i$  in rate based, rather than waiting time, terminology) and  $v_{g,i}$  is the NSE 'wait' time, i.e. 1 over the NSE rate ( $1/b_i$  in rate based terminology), and  $\sigma(i-1)$  is the probability a ribosome that initiates translation will reach the *i*th codon. Note that because the waiting time to a NSE,  $v_{g,i}$ , is so much greater than the elongation waiting time  $w_{g,i}$  we can ignore the actual variation in  $v_{g,i}$  between codons of the same type (this is easier to understand if you consider  $0 < b_i \ll 1$ ). Thus while we allow  $v_{g,i}$  to be codon specific, we treat each of these values as fixed and use  $v_{g,i}$  instead of  $v_{g,i}$ . Further, again because  $v_{g,i} \gg w_{g,i}$ , we can approximate  $\frac{w_{g,i} v_{g,i}}{w_{g,i} + v_{g,i}}$  as  $w_{g,i}$  based on a Taylor series expansion around  $1/v_{g,i} = 0$ . Thus,

$$p_{g,i}|w_{g,i} \propto \kappa_g \sigma(i-1) w_{g,i} \quad (9)$$

which is equivalent to the simple pausing time calculation except  $p_{g,i}$  is reduced by  $\sigma(i-1)$ .

The function  $\sigma(i-1)$  depends on the probability of successful elongation at the *i* - 1 upstream codons. When the waiting time for elongation each position *j* is known, then

$$\text{Pr}(\text{Elongation at position } j) = \frac{v_{g,j}}{w_{g,j} + v_{g,j}} \quad (10)$$

and

$$\sigma(i-1) = \prod_{j=1}^{i-1} \frac{v_{g,j}}{w_{g,j} + v_{g,j}} \quad (11)$$

However, when fitting the model we don't wish to actually estimate individual  $w_{g,j}$  values, but instead the parameters of the Gamma ( $\alpha_j, \lambda_j$ ) distribution the  $w_{g,j}$  are drawn from. Thus we treat  $w_{g,j}$  as a random variable whose uncertainty we integrate across to get an expected value. Letting  $f(\alpha, \lambda)$  represent the PDF

of the Gamma distribution, the expectation of the codon specific elongation probability is

$$E [\text{Pr}(\text{Elongation at position } j)] = \int_0^\infty \frac{v_{g,j}}{w_{g,j} + v_{g,j}} f(w_{g,j} | \alpha_j, \lambda_j) dw_{g,j} \quad (12)$$

$$= \lambda_j v_{g,j} \exp [\lambda_j v_{g,j}] E_{p=\alpha_j} (\lambda_j v_{g,j}) \quad (13)$$

where  $E_p(z)$  is the generalized exponential integral function (also referred to as Schlomilch functions (Oldham et al., 2009, p.380) and is represented as,

$$E_p(z) = \int_1^\infty \frac{e^{-zt}}{t^p} = z^{p-1} \int_z^\infty \frac{e^{-t}}{t^p} dt = z^{p-1} \Gamma(1-p, z) \quad (14)$$

where  $\Gamma(1-p, z)$  is the upper incomplete gamma function. See <http://dlmf.nist.gov/8.19> for more details. Note that, although the complete Gamma function is discontinuous at negative interger values, the upper incomplete is well defined even at negative integers so long as  $z > 0$ .

Using this formulation,

$$E [\text{Pr}(\text{Elongation at position } j)] = \exp [\lambda_j v_{g,j}] (\lambda_j v_{g,j})^{\alpha_j} \Gamma(1 - \alpha_j, \lambda_j v_{g,j}) \quad (15)$$

Assuming independence in  $w_{g,i}$  between positions means that the expected value of  $\sigma_g(i)$  is,

$$E [\sigma_g(i)] = \exp \left[ \sum_{j=1}^i \lambda_j v_{g,j} \right] \prod_{j=1}^i (\lambda_j v_{g,j} E_{\alpha_j} (\lambda_j v_{g,j})) \quad (16)$$

Note that these  $\lambda_j$  terms are equivalent to  $\lambda_c$  (but not  $\lambda'_c$ , which is used below).

To approximate  $E [\sigma_g(i)]$ , we set  $v_{g,j} = 1/(c_j b)$  where  $c_j$  is a codon specific scaling factor  $> 0$  and  $b$  is the mean nonsense error rate and taking a first order Taylor Series approximation around  $b = 0$ .

$$E [\sigma_g(i)] \approx 1 - \sum_{j=1}^i \frac{\alpha_j}{\lambda_j v_{g,j}} + O(1/v) \quad (17)$$

Noting that the reasoning which led to equations 2 and 3 for the pausing time model should still apply

if we use Equation (8) for  $p_{g,i}$ , rather than Equation (1). As a result,

$$\Pr(Y_{g,i}|\alpha_c, \lambda'_c, \psi_g, E[\sigma_g(i-1)]) = \frac{\Gamma(\alpha_c + Y_{g,i})}{\Gamma(\alpha_c) Y_{g,i}!} \left( \frac{\psi_g E[\sigma_g(i-1)]}{\lambda'_c + \psi_g E[\sigma_g(i-1)]} \right)^{Y_{g,i}} \left( 1 - \frac{\psi_g E[\sigma_g(i-1)]}{\lambda'_c + \psi_g E[\sigma_g(i-1)]} \right)^{\alpha_c} \quad (18)$$

where  $n_g$  is the number of amino acids encoded in gene  $g$ , i.e.  $n_g = \sum_c n_g^c$ . As before,  $\kappa_g \times m_g = \psi_g$  and  $\lambda'_c = \lambda_c Z/Y$ . Note also that  $Y_{g,i}!$  can be dropped from any Likelihood calculations. Thus, in this model we have both  $\lambda_c$  and  $\lambda'_c$ , thus we have an additional, genome wide parameter to estimate  $U = Z/Y$ . For completeness, we define  $\phi_g$  as the target production rate of the functionality produced by a complete, error free protein, i.e.  $\phi_g = \psi_g \sigma_g(n_g)$ . Thus, in the absence of translation errors,  $\phi_g$  is equal to the total synthesis rate protein  $g$ ,  $\psi_g$ . In contrast, in the presence of translation errors,  $\psi_g > \phi_g$  since the translation initiation rate must be elevated to compensate for the reduction in protein functionality due to translation errors. Furthermore, while the NSE model requires the likelihood for each position be calculated separately, the underlying terms for  $E[\sigma_g(i-1)]$  need only be calculated once per parameter evaluation since it is only how these terms are combined that varies between codons at different positions.

## Simulation

Unlike with the Pausing Time Model, we cannot aggregate codon specific RPF counts within a gene. Instead we have to simulate each position and as before we simulate in two steps. First, we generate  $w_{g,i}$  from  $\text{Gamma}(\alpha_c, \lambda'_c)$  and then we generate  $Y_{g,i}$  from  $\text{Poisson}(\psi_g \sigma_g(i-1) w_{g,i})$ . Note we are using  $\sigma_g(i)$  as defined in Eq. (11) and not  $E[\sigma_g(i)]$ , because when simulating data,  $w_{g,i}$  is known and, as a result, integration is unnecessary. Note that  $\psi_g$  values are scaled by the footprinting sequencing effort. Thus, we may find that you need scale all of your  $\psi_g$  values up or down by a constant term such that the expected number of footprints is on par with the number observed for a given set of genes.

## Parameter Definitions

	Definition	Units
$w_{g,i}$	Ribosome waiting/pausing/dwell time at codon position $i$ in gene $g$	1/t
$\alpha_c, \lambda_c$	Shape and rate parameter for distribution of waiting times for codon $c$ . The rate parameter is inversely related to average wait time, i.e. for codon $c$ $E[w_{g,i}] = \alpha_c/\lambda_c$	-
$m_g$	Density of mRNA transcripts for gene $g$ in cytosol.	1/Vol
$M_g$	Observed mRNA counts from RNA-Seq data.	1/Vol
$\kappa_g$	Rate constant determining ribosome initiation rate per mRNA. Function of diffusion of ribosomes, mRNA, and other factors.	$\frac{1}{\text{rib. mRNA Volt}}$
$\phi_g$	Target protein functionality production rate under a given set of conditions. Equal to $m_g\kappa_g\sigma_g(n_g)$ which is initiation rate discounted by the expected functionality of a protein.	1/t
$\psi_g$	Rate of initiation of protein synthesis under a given set of conditions. Equal to $m_g\kappa_g = \phi_g/\sigma_g(n_g)$ and, thus, is always greater than or equal to $\phi_g$ .	1/t
$\sigma_g(i)$	Probability ribosome reaches and successfully translate codon $i$ .	
$E[\sigma_g(i)]$	Expectation of $\sigma_g(i)$ used when fitting model to data.	
$p_{g,i}$	Probability a ribosome is found at position $i$ on an mRNA transcript from gene $g$ when translation initiation and completion of mRNA is at steady state.	
$P_{g,i}$	Probability of observing a footprint for position $i$ of mRNA from gene $g$ .	
$P_g^c$	Probability of observing a footprint for codon $c$ of mRNA from gene $g$ .	
$Z$	Partition function which scales the codon footprint sampling probabilities $P_{g,i}$ summed across $i$ and $g$ equals 1.	
$n_g^c$	Number of codons type $c$ in mRNA of gene $g$ .	
$Y_{g,i}$	Number of RFP observed for position $i$ in gene $g$	
$Y_g^c$	Number of RFP observed for codon $c$ in gene $g$	
$Y$	Total number of RFP in dataset.	
$\lambda'_c$	Composit parameter equal to $\lambda_c Z/Y$	
$\pi_j$	Prior probability for parameter $j$ .	

Table 1: Table of model parameters

## References

- C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, Hoboken, NJ, 4th edition, 2011.
- M. A. Gilchrist and A. Wagner. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *Journal of Theoretical Biology*, 239:417–434, 2006.
- K. Oldham, J. Myland, and J. Spanier. *An Atlas of Functions*. Springer, New York, 2nd edition, 2009.
- C. Pop, S. Rouskin, N. T. Ingolia, L. Han, E. M. Phizicky, J. S. Weissman, and D. Koller. Causal signals between codon bias, mrna structure, and the efficiency of translation and elongation. *Molecular Systems*

*Biology*, 10(12):770, 2014. ISSN 1744-4292. doi: 10.15252/msb.20145524. URL <http://dx.doi.org/10.15252/msb.20145524>. 770.