History

- Initial model by mikeg on 6/10/15.
- NSE model added by mikeg on 7/21/15.
- Compiled on Wednesday 29$^{\text{th}}$ July, 2015at 13:21

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 κ_g is a rate constant scales the average rate at which ribsomes 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 ribsome 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 α_c is the shape parameter, λ_c is the scale 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.

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

$$P_{g,i} = p_{g,i} m_g / Z \tag{2}$$

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,

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

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

Letting n_s be the footprint sample size, i.e. the number of footprints examined, where $n_s \gg 1$ 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 $n_s P_{g,i}$. 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 $\lambda'_c = \lambda_c Z / n_s$. Explicitly,

$$\Pr\left(Y_{g,i} | \alpha_c, \lambda'_c, m_g, \kappa_g\right) = \frac{\Gamma\left(\alpha_c + Y_{g,i}\right)}{\Gamma\left(\alpha_c\right) Y_{g,i}!} \left(\frac{m_g \kappa_g}{\lambda'_c + m_g \kappa_g} \right)^{Y_{g,i}} \left(\frac{\lambda'_c}{\lambda'_c + m_g \kappa_g} \right)^{\alpha_c} \tag{4}$$

Note that m_g and κ_g are gene g and environment specific terms which can be equated to the equilibrium protein synthesis rate ϕ_g for gene g under the experimental conditions, i.e. $\phi_g = \kappa_g m_g$. The composite parameter λ'_c consists of the codon specific scale term λ_c and the ratio of Z to n_s, two genome wide parameters.

Given the properties of the NB 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,

$$\mathrm{L}\left(\phi_g, \alpha_c, \lambda'_c \middle| Y_g^c, n_g^c\right) = \frac{\Gamma\left(n_g^c \alpha_c + Y_g^c\right)}{\Gamma\left(n_g^c \alpha_c\right) Y_g^c!} \left(\frac{\phi_g}{\lambda'_c + \phi_g}\right)^{Y_g^c} \left(\frac{\lambda'_c}{\lambda'_c + \phi_g}\right)^{n_g^c \alpha_c} \tag{5}$$

$$= \frac{\Gamma\left(\alpha'_c + Y_g^c\right)}{\Gamma\left(\alpha'_c\right) Y_g^c!} \left(\frac{\phi_g}{\lambda'_c + \phi_g}\right)^{Y_g^c} \left(\frac{\lambda'_c}{\lambda'_c + \phi_g}\right)^{\alpha'_c} \tag{6}$$

where n_g^c is the number of times codon c is found in the ORF of gene g and $\alpha'_c = n_g^c \times \alpha_c$. The total Likelihood of the data is

$$\mathrm{L}\left(\vec{\phi_g}, \vec{\alpha}_c, \vec{\lambda'_c} \middle| \vec{Y_g^c}, \vec{n_g^c}\right) = \prod_{g \in \mathbb{G}} \prod_{c \in \mathbb{C}} \mathrm{L}\left(\phi_g, \alpha_c, \lambda'_c \middle| Y_g^c, n_g^c\right) \tag{7}$$

Note that using RFP data alone, $\kappa_g \times m_g = \phi$ and $\lambda_c Z$ 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.) 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 L from Gamma(α_c, λ'_c) or Gamma(α'_c, λ'_c), then pull Y from Poisson(L).

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

Simulation

Although our Likelihood function can be described using a Negative Binomial distribution, because $x = \alpha$ and α is not discrete, we cannot use standard NB routines to simulate our data. Instead we need simulate in two steps. First, we generate W_c from Gamma(α'_c, λ'_c) and then we generate Y_g^c from Poisson $(\phi_g W_c)$

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,

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

Where, as before, κ_g is the translation initiation rate constant per mRNA$_g$, $w_{g,i}$ 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 ith 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. 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,

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

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. Letting $f(\alpha, \lambda)$ represent the PDF of the Gamma distribution, the codon specific elongation probability is

$$\Pr(\text{Elongation at position } j) = \int_0^\infty \frac{v_{g,j}}{w_{g,j}+v_{g,j}} f\left(w_{g,j}|\alpha_j, \lambda_j\right) dw_{g,j} \tag{10}$$

$$= \exp\left[\lambda_j v_{g,j}\right] E_{n=\alpha_j}\left(\lambda_j v_{g,j}\right) \tag{11}$$

where $E_n(x)$ is the exponential integral function. Assuming independence in $w_{g,i}$ between positions,

$$\sigma(i) = \exp\left[\sum_{j=1}^{i} \lambda_j v_{g,j}\right] \prod_{j=1}^{i} E_{\alpha_j}\left(\lambda_j v_{g,j}\right) \tag{12}$$

Note that these λ_j terms are equivalent as λ_c, but not λ'_c.

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,

$$\Pr\left(Y_{g,i}|\alpha_i, \lambda'_i, m_g, \kappa_g, \sigma(i-i)\right) = \frac{\Gamma\left(\alpha_i + Y_{g,i}\right)}{\Gamma\left(\alpha_i\right) Y_{g,i}!} \left(\frac{m_g \kappa_g \sigma(i-1)}{\lambda'_i + m_g \kappa_g \sigma(i-1)}\right)^{Y_{g,i}} \left(\frac{\lambda'_i}{\lambda'_i + m_g \kappa_g \sigma(i-1)}\right)^{\alpha_i} \tag{13}$$

While the NSE model requires the likelihood for each position be calculated separately, the underlying terms for $\sigma(i-1)$ 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_i from Gamma(α_i, λ_i') and then we generate $Y_{g,i}$ from Poisson $(\phi_g \sigma(i-1) W_i)$

Parameter Definitions

References

Gilchrist, M. A. and Wagner, A. (2006). A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *Journal of Theoretical Biology*, **239**, 417–434.

Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., and Koller, D. (2014). Causal signals between codon bias, mrna structure, and the efficiency of translation and elongation. *Molecular Systems Biology*, **10**(12), 770. 770.

	Definition	Units
$w_{g,i}$	Ribosome waiting/pausing/dwell time at codon position i in gene g	1/t
α_c, λ_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
κ_g	Rate constant determining ribosome initiation rate per mRNA. Function of diffusion of ribosomes, mRNA, and other factors.	$\frac{1}{\text{rib. mRNA Vol}t}$
ϕ_g	Steady state protein production rate. Equal to $m_g \times \kappa_g$.	
$p_{g,i}$	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.	
n_s	RPF sample size. That is, number of fragments in dataset.	
$Y_{g,i}$	Number of RFP observed for position i in gene g	
Y_g^c	Number of RFP observed for codon c in gene g	
λ_c'	Composit parameter equal to $\lambda_c Z/n_s$	
π_j	Prior probability for parameter j.	

Table 1: Table of model parameters