Description of the mixture model implemented in the R package ribModel

Cedric Landerer

December 21, 2015

1 General

Table 1: Variable legend

Variable	**Description**
G	Number of genes
g	Gene index
A	Number of Amino Acids
a	Amino acid index
C	Number of Codons
c	Codon index
$\vec{\zeta}$	Vector of codon counts
ζ_c	Codon count for codon c
f	log Likelihood function
Θ	Set of parameters for f
I	Number of ϕ obs.
ι	ϕ obs. index
Φ	Obs. ϕ value
ϕ	estm. ϕ value
P	Posterior
Υ	Posterior Jacobian adjusted
z	Mixture
Z	Number of mixtures

Table 2: Parameter distributions

Parameter	**Distribution**	**proposal**
ϕ	$LN(-\frac{s_\phi^2}{2}, s_\phi)$	$\log(\phi') \sim N(\log(\phi), \sigma_\phi)$
s_ϕ	$U(0, \infty)$	$\log(s'_\phi) \sim N(\log(s_\phi), \sigma_{s_\phi})$
ΔM	$N(0, \sigma)$	$\Delta M' \sim N(\Delta M, \sigma)$
$\Delta\eta$	$U(-\infty, \infty)$	$\Delta\eta' \sim N(\Delta\eta, \sigma)$
p_z	$Dir(\alpha)$	Gibbs sampled
A_ϕ	$U(-\infty, \infty)$	$A'_\phi \sim N(A_\phi, \sigma_{A_\phi})$
s_ϵ	$G^{-1}(\frac{G-1}{2}, \frac{S^2}{2})$	Gibbs sampled

1.1 Calculating the Posterior trace

We calculate the unscaled log posterior as shown in equation 1 as

$$P(\phi|\zeta, z, \Delta M, \Delta\eta, s_\phi, A_\phi, s_\epsilon) \propto \sum_g^G \sum_z^Z \left(p_{z,g} P_g(\phi_{z,g}|\vec{\zeta}, \Theta_z) \right) + \sum_a^A \log(\pi(\Delta M_a|0, \sigma)) \tag{1}$$

where $p_{z,g}$ is the probability of gene g being assigned to mixture z (see eqn 2), and $\Upsilon_g(\phi_{z,g}|\zeta, \Theta_z)$ is the log posterior for gene g under the parameters defining mixture z (see eqn 7). We assume ΔM to be normal distributed with $\pi(\Delta M_a|0, \sigma)$, ($\sigma = x \; \forall a$), and take it as prior into account when calculating the unscaled log posterior.

1.2 Calculate mixture assignment

Genes are assigned to mixtures based on there likelihood calculated under each mixture weighted by the probability of that mixture. The probability for a gene being in each mixture is calculated as shown in eqn. 2 and used to draw the mixture assignment from a multinomial distribution $z_g \sim \text{Multinomial}(Z, p_{1,g} \dots p_{Z,g})$

$$p_{z,g} = \exp\left(\frac{\log(p_z) + P_g(\phi_{z,g}|\vec{\zeta}, \Theta_z)}{\sum_z^Z \log(p_z) + P_g(\phi_{z,g}|\vec{\zeta}, \Theta_z)} \right) \tag{2}$$

2 ROC

We calculate the probability for each synonymous codon coding for an amino acid using equation 3.

$$p_{g,c} = \frac{-\Delta M_c - \Delta\eta_c \phi_g}{\sum_c^C -\Delta M_c - \Delta\eta_c \phi_g} \tag{3}$$

where ΔM is the difference in mutation bias between the current codon and a chosen reference codon, and $\Delta\eta$ is the difference in translation efficiency.

We assume codon count being distributed according to a multinomial distribution where the probability of each synonymous codon is given by equation 3. The log likelihood for a set of parameters describing each amino acid is therefore given by equation 4

$$f_a(\cdot|\zeta, \Theta_z) = \sum_c^C \log(p_{g,c})\zeta_c \tag{4}$$

2.1 Accept/Reject ϕ

To accept/reject a proposed ϕ value, we calculate the log posterior as follows. The likelihood function is the sum over all likelihoods for the amino acids (eqn 4) in the gene of interest shown in equation 5.

$$f_g(\phi_g|\zeta, \Theta_z) = \sum_a^A f_a(\phi_g|\vec{\zeta}, \Theta_z) \tag{5}$$

A log normal prior ($p(\phi_g| - \frac{s_\phi^2}{2}, s_\phi)$) on ϕ is assumed.

$$P_g(\phi_g|\zeta, \Theta_z) = f_g(\phi_g|\zeta, \Theta_z) + \log(p(\phi_g| - \frac{s_\phi^2}{2}, s_\phi)) + \sum_\iota^I \log(p(\log \Phi_{g,\iota}| \log \phi_g + A_{\phi_\iota}, s_{\epsilon\iota})) \tag{6}$$

All terms $\log(p(\log \Phi_{g,\iota} + A_{\phi_\iota}| \log \phi_g, s_{\epsilon_\iota}))$ are 0 if no observed ϕ values Φ are available.

Since we assume the ϕ values to be log normal distributed, we take the reverse jumb probability into account in equation 7.

$$\Upsilon_g(\phi_g|\zeta, \Theta_z) = P_g(\phi_g|\zeta, \Theta_z) - \log(\phi_g) \tag{7}$$

The ratio of two log normal distributions can be simplified to the ratio $\frac{\phi}{\phi'}$ which results in the term $\log \phi$ on the log scale.

The acceptance/rejection of a proposed ϕ_g' is performed on the logscale and described in equation 8 where $r \sim Exp(1)$ and α is given by equation 9

$$\phi_g = \begin{cases} \phi_g', & \text{if } -r < \alpha \\ \phi_g, & \text{else} \end{cases} \tag{8}$$

$$\alpha = \Upsilon_g(\phi_g'|\zeta, \Theta_z) - \Upsilon_g(\phi_g|\zeta, \Theta_z) \tag{9}$$

2.2 Accept/Reject s_ϕ

Since we are not interested in the likelihood for s_ϕ, we calculate the likelihood ratio necessary for the acceptance/rejection step directly. Since s_ϕ has to be positive, we propose a new $\log s_\phi$ using a normal distributed random walk. Therefore it follows that s_ϕ is drawn from a log-normal distribution. This forces us to take the reverse jump probability into account since the log-normal distribution is not symmetric. The ratio can be simplified since symmetric elements will cancel. We take the reverse jumb probability ratio for each mixture element into accountas as shown in equation 10.

$$\Delta J = \frac{\log(J)}{\log(J')} = \sum_{z}^{Z} -(\log(s_{\phi_z}) - \log(s'_{\phi_z})) \tag{10}$$

We then calculate the probability ratio of the data (here the current ϕ values estimated in the previous step) given the current and proposed s_ϕ values as shown in equation 11.

$$\Delta f(s_\phi|\phi_z) = \sum_{g}^{G} \left(\log(LN(\phi_{g,z}| - \frac{s_{\phi_z}^2}{2}, s_{\phi_z})) - \log(LN(\phi_{g,z}| - \frac{s'^{\,2}_{\phi_z}}{2}, s'_{\phi_z})) \right) \tag{11}$$

The reverse jump probability ratio and the likelihood ratio are combined on the log scale to obtain the log acceptance ratio α as shown in equation 12.

$$\alpha = \Delta f_z(s_\phi|\phi_z) + \Delta J \tag{12}$$

The acceptance/rejection of proposed s'_ϕ is given by equation 13 where $r \sim Exp(1)$ and α is given by equation 12.

$$\phi_g = \begin{cases} s'_\phi, & \text{if } -r < \alpha \\ s_\phi, & \text{else} \end{cases} \tag{13}$$

2.3 Accept/Reject ΔM and $\Delta\eta$

The codon specific parameter ΔM and $\Delta\eta$ are acceptted/rejected together for each amino acid. A new set of $\Delta M'$ and $\Delta\eta'$ values θ, is proposed from a normal distribution $\theta' \sim N(\theta, \Sigma)$, where Σ is a covariance matrix where only the variance terms on the diagonal are non zero. We assume as prior term a normal distribution form ΔM and a uniform prior for $\Delta\eta$. The posterior is for a set of ΔM and $\Delta\eta$, θ is calculated as

$$P_a(\Delta M_a, \Delta\eta_a|\zeta, \phi, z) = \sum_{g}^{G} f_a(\Delta M_a, \Delta\eta_a|\zeta, \phi_{g,z}) + \log(\pi(\Delta M_a|0, \sigma)) \tag{14}$$

where $\pi(\Delta M_a|0, \sigma)$ is the normal distributed prior on ΔM.

The acceptance/rejection of a proposed set of ΔM_a and $\Delta\eta_A$ is given by equation 15 where $r \sim Exp(1)$ and α is given by equation 16

$$\phi_g = \begin{cases} \Delta M'_a, \Delta\eta'_a, & \text{if } -r < \alpha \\ \Delta M_a, \Delta\eta_a, & \text{else} \end{cases} \tag{15}$$

$$\alpha = P_a(\Delta M'_a, \Delta\eta'_a|\vec{\zeta}, \vec{\phi_z}) - P_a(\Delta M_a, \Delta\eta_a|\vec{\zeta}, \vec{\phi_z}) \tag{16}$$

3 FONSE