REU 2013 Summer Student Analysis

Multinomial Parameter Analysis

This was primarily written by Christopher Oballe.

Assuming that the probability of finding a specific synonymous codon in gene sequence $\vec{x}$ given expression level ϕ and position i follows a multinomial logisitic regression, we have:

$$\Pr(\vec{x}|\phi, i) \propto \prod_{i=1}^{n} \exp\left(\theta_0(x_i) + \theta_y(x_i)y + \theta_{yi}(x_i)yi\right) \tag{214}$$

Here, x_i denotes the codon at position i, θ_k denotes the parameter associated with independent variable k, and y denotes the aggregate variable $-N_e q\phi$.

Alternatively, from Shah and Gilchrist (2011), we have:

$$\Pr(\vec{x}|\phi) \propto \exp\left(\sum \ln(\mu_i) \ + y\eta\right) \tag{215}$$

where μ_i denotes the mutation rate of codon type i

Using the first order approximation for η in 215 yields:

$$\Pr(\vec{x}|\phi) \propto \exp\left(\sum_{i=1}^{n} \ln(\mu_i(x_i)) + y\sum_{i=1}^{n} \ \omega(x_i)\beta_i) = \prod_{i=1}^{n} \exp(\ln(\mu_i(x_i)) + y\omega(x_i)\beta_i\right) \tag{216}$$

$$= \prod_{i=1}^{n} \exp\left(\ln(\mu_i(x_i)) + y\omega(x_i)(a_1 + a_2(i-1))\right) \tag{217}$$

$$= \prod_{i=1}^{n} \exp\left(\ln(\mu_i(x_i)) + y\omega(x_i)(a_1 - a_2) + yi\omega(x_i)a_2\right) \tag{218}$$

Comparing 218 with 214 reveals the identities of the parameters in the multinomial logisitic regression. For any given synonymous codon, the intercept parameter is the log of its mutation rate. The expression level parameter is the product of the odds ratio of failed elongation to successful elongation, $-qN_e$, as well as the difference in overhead protein production cost and elongation cost. Finally, the the position:expression level parameter is the product of the odds ratio, elongation cost, and $-qN_e$.

Alternative Approximation to Eta

Recall the definition of η from 3. Our strategy is to develop an approximation for all higher order deriviatives of η with respect to b, evaluate these approximations at $b = 0$, and use these results to build a Taylor series. Since η is a polynomial of degree n, it's actually equal to the first n terms of its Taylor series expansion, so our approximation for the η Taylor series can be more simply viewed as an approximation to η itself.

η is a sum of terms, so its derivatives are equal to the sums of the derivatives of its terms, and $a_1 + a_2 n$ is constant with respect to b and consequently evaluates to zero upon differentiation. Therefore, it's sufficient to find the derivatives of $(a_1 + a_2 i)\frac{b}{c_i}\prod_{k=i+1}^{n}(1 + \frac{b}{c_k})$ then add them all together in order to find a derivative for η.

Since deriving $(a_1 + a_2 i)\frac{b}{c_i}\prod_{k=i+1}^{n}(1 + \frac{b}{c_k})$ will involve taking the derivaitive of a complicated product,

it's worthwhile to state a formula for differentiating long products.

$$\frac{d}{dx}\prod_{i=1}^{m} f_i(x) \tag{219}$$

$$= \sum_{i=1}^{m}\left(\frac{d}{dx}f_i(x)\right)\prod_{j\neq i} f_j(x) \tag{220}$$

$$= \prod_{j=1}^{m} f_j(x)\sum_{i=1}^{m}\frac{\frac{d}{dx}f_i(x)}{f_i(x)} \tag{221}$$

$$= \prod_{j=1}^{m} f_j(x)\frac{d}{dx}\sum_{i=1}^{m}\ln\left(f_i(x)\right) \tag{222}$$

220 follows from the product rule and a basic induction argument. 221 and 222 were obtained purely through algebraic manipulation.

Using 222, we'll directly compute the first derivative to gain some insight.

$$\frac{d}{db}\left((a_1+a_2i)\frac{b}{c_i}\prod_{k=i+1}^{n}(1+\frac{b}{c_k})\right) \tag{223}$$

$$= (a_1+a_2i)\frac{1}{c_i}\prod_{k=i+1}^{n}(1+\frac{b}{c_k}) + (a_1+a_2i)\frac{b}{c_i}\prod_{k=i+1}^{n}(1+\frac{b}{c_k})\frac{d}{db}\sum_{j=i+1}^{n}\ln(1+\frac{b}{c_j}) \tag{224}$$

$$\approx (a_1+a_2i)\frac{1}{c_i}\prod_{k=i+1}^{n}(1+\frac{b}{c_k}) + (a_1+a_2i)\frac{b}{c_i}\prod_{k=i+1}^{n}(1+\frac{b}{c_k})\sum_{j=i+1}^{n}\frac{1}{c_j} \tag{225}$$

The approximation follows from the identity $\ln(1+x)\approx x$ for $x << 1$. Alternatively, one can reach the same result using 221 :

$$\frac{d}{db}\left(\prod_{k=i+1}^{n}(1+\frac{b}{c_k})\right) \tag{226}$$

$$= \prod_{k=i+1}^{n}(1+\frac{b}{c_k})\sum_{j=i+1}^{n}\frac{\frac{d}{db}(1+\frac{b}{c_j})}{1+\frac{b}{c_j}} \tag{227}$$

$$= \prod_{k=i+1}^{n}(1+\frac{b}{c_k})\sum_{j=i+1}^{n}\frac{\frac{1}{c_j}}{1+\frac{b}{c_j}} \tag{228}$$

$$\approx \prod_{k=i+1}^{n}(1+\frac{b}{c_k})\sum_{j=i+1}^{n}\frac{1}{c_j} \tag{229}$$

Here, the approximation follows from the fact that $\frac{b}{c_j}\approx 0$ for all j. Applying this approximation successively, it's easy to see that the mth derivative of $\left(\prod_{k=i+1}^{n}(1+\frac{b}{c_k})\right)$ is $\left(\sum_{j=i+1}^{n}\frac{1}{c_j}\right)^m\left(\prod_{k=i+1}^{n}(1+\frac{b}{c_k})\right)$.

To make further computations more compact, let $\beta_i = a_1 + a_2 i$. Computing the second derivative yields:[3]

$$\frac{d^2}{db^2}\left(\beta_i \frac{b}{c_i} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k})\right) \tag{230}$$

$$\approx \beta_i \frac{1}{c_i} \sum_{j=i+1}^{n} \frac{1}{c_j} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) + \beta_i \frac{1}{c_i} \sum_{j=i+1}^{n} \frac{1}{c_j} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) + \beta_i \frac{b}{c_i} \left(\sum_{j=i+1}^{n} \frac{1}{c_j}\right)^2 \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) \tag{231}$$

$$= 2\beta_i \frac{1}{c_i} \sum_{j=i+1}^{n} \frac{1}{c_j} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) + \beta_i \frac{b}{c_i} \left(\sum_{j=i+1}^{n} \frac{1}{c_j}\right)^2 \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) \tag{232}$$

Note that at b=0, $\frac{d}{db}\beta_i \frac{b}{c_i} \prod_{k=i+1}^{n}(1 + \frac{b}{c_k}) \approx \beta_i \frac{1}{c_i}$ and $\frac{d^2}{db^2}\beta_i \frac{b}{c_i} \prod_{k=i+1}^{n}(1 + \frac{b}{c_k}) \approx 2\beta_i \frac{1}{c_i} \sum_{j=i+1}^{n} \frac{1}{c_j}$.

We'll now show by induction that under our derivative approximation, the m^th derivative of $\beta_i \frac{b}{c_i} \prod_{k=i+1}^{n}(1 + \frac{b}{c_k})$ evaluated at 0 is equal to:

$$m\beta_i \frac{1}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m-1} \tag{233}$$

We've already shown that our hypothesis holds for 1 and 2. Assume that for an arbitrary natural number m, the m^th derivative equals

$$m\beta_i \frac{1}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m-1} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) + \beta_i \frac{b}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) \tag{234}$$

Applying the derivative approximation for the $m + 1$ case results in:

$$\frac{d}{db}\left[m\beta_i \frac{1}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m-1} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) + \beta_i \frac{b}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k})\right] \tag{235}$$

$$\approx m\beta_i \frac{1}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m} \prod_{k=i+1}^{n} \left(1 + \frac{b}{c_k}\right) + \beta_i \frac{1}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m} \prod_{k=i+1}^{n} \left(1 + \frac{b}{c_k}\right) + \beta_i \frac{b}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m+1} \prod_{k=i+1}^{n} (1 + \frac{b}{c_k}) \tag{236}$$

$$= (m+1)\beta_i \frac{1}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m} \prod_{k=i+1}^{n} \left(1 + \frac{b}{c_k}\right) + \beta_i \frac{b}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m+1} \prod_{k=i+1}^{n} \left(1 + \frac{b}{c_k}\right) \tag{237}$$

Observe at $b = 0$ $\frac{d^{m+1}}{db^{m+1}}\beta_i \frac{b}{c_i} \prod_{k=i+1}^{n}(1 + \frac{b}{c_k}) \approx (m+1)\beta_i \frac{1}{c_i} \left[\sum_{j=i+1}^{n} \frac{1}{c_j}\right]^{m}$ to conclude the proof.

Now we can construct an approximation to any term η_i of η:

$$\eta_i \approx \frac{\beta_i}{c_i} b + \frac{2\frac{\beta_i}{c_i}}{2!} \sum_{k=i+1}^{n} \frac{1}{c_k} b^2 + ... + \frac{(n)\frac{\beta_i}{c_i}}{(n!)} \left[\sum_{k=i+1}^{n} \frac{1}{c_k}\right]^{n-1} b^n \tag{238}$$

[3] mikeg: How dependent is this on the exact form of β_i?

Letting ω_i denote the odds ratio $\frac{b}{c_i}$ and $\omega_{i+1,n}$ denote the sum of the terminal odds ratios $\sum_{j=i+1}^{n} \frac{1}{c_j}$, we can rewrite 238 as

$$\begin{aligned}\eta_t &\approx \beta_i \omega_i \left(1 + \omega_{i+1,n} + \frac{1}{2!}\omega_{i+1,n}^2 + ... + \frac{1}{n-1!}\omega_{i+1,n}^{n-1}\right) \\ &\approx \beta_i \omega_i e^{\omega_{i+1,n}}\end{aligned}$$

Thus,

$$\eta \approx a_1 + a_2 n + \sum_{i=1}^{n} \beta_i \omega_i e^{\omega_{i+1,n}} \tag{239}$$

Jeremy Rogers (Undergrad Researcher 2015-2016) Notes on FONSE Model

From the REU 2013 paper, to build our model, we began with an equation from population genetics describing the probability of fixation for a gene sequence, represented as a vector of codons $\vec{c}$, given gene expression level ϕ, given by

$$\Pr(\vec{c}) \propto \exp\left(-q2N_e\phi\eta(\vec{c}) + \sum_{i=1}^{n} M_i\right) \tag{240}$$

In the above equation, i denotes position n is the length of the gene $\vec{c}$, q is a scaling constant, N_e is the effective population size, M_i is the mutation bias term for the codon at position i relative to its synonyms, and $\eta(\vec{c})$ is the expected cost of protein production. To make future equations more concise, y_1 will denote the aggregate variable $-q2N_e\phi$. Intuitively, this equation is stating that one is less likely to observe synonymous gene sequences that produce polypeptides expensively and that this effect is more apparent for highly expressed genes in large populations. Our model uses a definition of η based on average nonsense error cost.

$$\eta(\vec{c}) = (a_1 + a_2 n) + \sum_{i=1}^{n}(a_1 + a_2(i-1))\frac{p_i}{1-p_i}\prod_{k=i+1}^{n}(1 + \frac{p_k}{1-p_k}) \tag{241}$$

In the above equation, a_1 is the ribosome initiation cost of protein production, a_2 is the cost of peptide addition, and p_i is the probability of a nonsense error for the codon at position i. If we let ω_i denote the ratio of failed elongation to successful elongation $\frac{p_i}{1-p_i}$, then, as seen in Equation (89), the first order approximation of η around $p_i = 0$ is

$$\eta(\vec{c}) \approx (a_1 + a_2 n) + \sum_{i=1}^{n}\omega_i[a_1 + a_2(i-1)] \tag{242}$$

Substituting this result into the previous equation for $\Pr(\vec{c})$ yields

$$\Pr(\vec{c}) \propto \exp\left(y_1\left[a_1 + a_2 n + \sum_{i=1}^{n}(a_1 + a_2 i)\omega_i\right] + \sum_{i=1}^{n}\ln(\mu_i)\right) \tag{243}$$

$$= \exp(y_1[a_1 + a_2 n])\prod_{i=1}^{n}\exp[\ln\mu_i + \omega_i(a_1 - a_2)y_1 + \omega_i a_2 y_1 i] \tag{244}$$

For genes in a synonymous sequence space, the term appearing outside of the product will be constant, and since probabilities are found through scaling the equation by its sum over all genes in the synonymous sequence space, the preceding exponential term will cancel during calculations. Thus, it can be neglected for our purposes, allowing us to state it in a slightly more succinct form:

$$\Pr(\vec{c}|\phi) \propto \prod_{i=1}^{n}\exp[\ln\mu_i + \omega_i(a_1 - a_2)y_1 + \omega_i a_2 y_1 i] \tag{245}$$

Inspection reveals that the probability of observing $\vec{c}$ is simply the product of the probabilities of observing each of its c_i at position i, where the latter probabilities are proportional to $\exp[\ln\mu_i + \omega_i(a_1 - a_2)y_1 + \omega_i a_2 y_1 i]$. Note also that the equation implies positional independence, i.e. the identity of a codon at position i does not affect the probability of observing another codon at position j for $i \neq j$. To find the exact probabilities of observing particular members of groups of synonyms, one would compute for each codon in the synonymous space and divide its value by the sum of all values. This forms the basis of the first order model.

$$\Pr(c_i|\phi, i) = \frac{\exp[\ln\mu_i + \omega_i(a_1 - a_2)y_1 + \omega_i a_2 y_1 i}{\sum_{u=1}^{m}\exp[\ln\mu_u + \omega_u(a_1 - a_2)y_1 + \omega_u a_2 y_1 i]} \tag{246}$$

Namely, our model predicts the probability of observing c_i amongst its m synonyms at position i in a gene with expression level ϕ. Note that our model assumes the probabilities of observing specific codons follow a multinomial regression with parameters $\ln \mu_i$, $\omega_i(a1 - a2)$, and $\omega_i a_2$, associated with the intercept, y_1, and $y_1 i$, respectively. Thus, we now have a means of estimating these biological parameters by fitting multinomial regressions to real genome data. It is also important to note that Shah and Gilchrist's ribosome overhead cost (ROC) model also assumes a multinomial regression, but its parameter identities are different. Specifically, the ROC model has $\ln \mu_i$ as the intercept parameter and elongation time as a parameter associated with y_1. It does not consider the y_1 - position interaction. Nonetheless, we can evaluate the accuracy of each models predictions to speculate about the relative contributions of nonsense errors and ribosome pausing times to codon usage bias.