Algorithm for inference of enrichment ratios

Jesse D. Bloom

May 29, 2013

We quantify the effect of a mutation in terms of its enrichment after selection for viral growth. This enrichment can be thought of as the frequency of a mutation in the virus mutant library (**mutvirus**) divided by its frequency in the initial plasmid mutant library (**mutDNA**), after correcting for sources of error as described below. Highly deleterious mutations will have enrichment ratios close to zero. Mutations that enhance viral growth or increase mutational tolerance will have enrichment ratios greater than one. Note that an enrichment ratio cannot simply be interpreted as the effect of a mutation on viral growth – because most clones in our libraries have multiple mutations, enrichment summarizes the effect of a mutation in both the wildtype gene and a variety of closely related mutants. A mutation can therefore change frequency due to its inherent effect on viral growth or its effect on the gene's ability to tolerate other mutations.

Estimating enrichment ratios simply by dividing the frequency in the **mutvirus** library by the frequency in the **mutDNA** library is problematic for several reasons. First, such an approach fails to account for sources of error (such as sequencing) that affect the observed frequencies of mutations. Second, the libraries contain finite numbers of counts for each mutation, and estimating ratios by dividing counts from finite samples is a notoriously statistically biased approach. For example, in the limiting case where a mutation is counted once in the **mutvirus** library and not at all in the **mutDNA** library, taking the ratio of counts gives an enrichment of infinity and suggests that that the mutation is extremely favorable – yet in practice such counts give us little confidence that we have reliably measured the true effect of the mutation.

To circumvent these problems, we use a Bayesian approach. We begin with prior estimates that every mutation has properties equal to the average for all mutations in the library. We specify likelihood functions that give the probability of observing a set of counts for a mutation given its frequency in the initial library, frequency of erroneous counts, and enrichment ratio. We then combine these priors and likelihood functions to estimate the posterior distributions of the enrichment ratios. This approach accounts for sources of experimental error and avoids overfitting enrichment ratios for mutations with low counts.

In implementing this approach, we use the counts in the **DNA** library to quantify errors due to PCR and and sequencing. We use the counts in the **RNA** library to quantify errors due to reverse-transcription. We assume that transcription of the viral genes from the reverse-genetics plasmids and subsequent replication of these genes by the influenza polymerase introduces a negligible number of new mutations relative to the number already present in the plasmid mutant library.

At each codon site r of the gene, there are 63 non-wildtype codon identities. Let i be one of these non-wildtype codons. Let $N_r^{\rm DNA}$ be the total number of sequencing reads at site r in the **DNA** library, and let $n_{r,i}^{\rm DNA}$ be the number of these reads that report a mutation of site r to codon i. Similarly, let $N_r^{\rm mutDNA}$, $N_r^{\rm RNA}$, and $N_r^{\rm mutvirus}$ be the total number of reads at site r and let $n_{r,i}^{\rm mutDNA}$, $n_{r,i}^{\rm RNA}$, and $n_{r,i}^{\rm mutvirus}$ be the total number of these reads that report a mutation of site r to codon i in the **mutDNA**, **RNA**, and **mutvirus**, respectively. Let $\epsilon_{r,i}$ be the rate at which site r is erroneously read to be codon i due to PCR or sequencing errors, such that $\epsilon_{r,i} = \lim_{N_r^{\rm DNA} \to \infty} \left(\frac{n_{r,i}^{\rm DNA}}{N_r^{\rm DNA}} \right)$. Let $\rho_{r,i}$ be the rate at which site r is erroneously copied be codon i during reverse-transcription, such that $\rho_{r,i} + \epsilon_{r,i} = \lim_{N_r^{\rm RNA} \to \infty} \left(\frac{n_{r,i}^{\rm RNA}}{N_r^{\rm RNA}} \right)$. Let $\mu_{r,i}$ be the rate at which site r is mutated to codon i in the plasmid mutant library, such that $\mu_{r,i} + \epsilon_{r,i} = \lim_{N_r^{\rm mutDNA} \to \infty} \left(\frac{n_{r,i}^{\rm mutDNA}}{N_r^{\rm mutDNA}} \right)$. Let $\phi_{r,i}$ be the enrichment during the viral growth of clones that contain the mutation of site r to i, such

that $\phi_{r,i} \times \mu_{r,i} + \rho_{r,i} + \epsilon_{r,i} = \lim_{N_r^{\text{mutvirus}} \to \infty} \left(\frac{n_{r,i}^{\text{mutvirus}}}{N_r^{\text{mutvirus}}} \right)$. We assume that the rates $\epsilon_{r,i}$, $\rho_{r,i}$, and $\mu_{r,i}$ are all $\ll 1$ and so neglect the possibility that a clone experiences experiences more than one of these sources of mutation at a single site.

If we assume that the vast majority of clones retain the wildtype identity at any given site, then we can neglect the correlations between the counts for different mutant codons i at a given site r. In this case, the probability of observing $n_{r,i}^{\text{DNA}}$ counts is given by a Poisson distribution with mean $N_r^{\text{DNA}} \times \epsilon_{r,i}$ and similar results hold for the other counts. Specifically, define

$$f\left(k;\lambda\right) = e^{-\lambda} \times \frac{\lambda^k}{k!} \tag{1}$$

to be the Poisson probability of observing k events when the expected number is λ. Then we have the following likelihood functions:

$$\Pr\left(n_{r,i}^{\text{DNA}} \mid N_r^{\text{DNA}}, \epsilon_{r,i}\right) = f\left(n_{r,i}^{\text{DNA}}; N_r^{\text{DNA}} \times \epsilon_{r,i}\right) \tag{2}$$

$$\Pr\left(n_{r,i}^{\text{RNA}} \mid N_r^{\text{RNA}}, \epsilon_{r,i}, \rho_{r,i}\right) = f\left(n_{r,i}^{\text{RNA}}; N_r^{\text{RNA}} \times [\epsilon_{r,i} + \rho_{r,i}]\right) \tag{3}$$

$$\Pr\left(n_{r,i}^{\text{mutDNA}} \mid N_r^{\text{mutDNA}}, \epsilon_{r,i}, \mu_{r,i}\right) = f\left(n_{r,i}^{\text{DNA}}; N_r^{\text{mutDNA}} \times [\epsilon_{r,i} + \mu_{r,i}]\right) \tag{4}$$

$$\Pr\left(n_{r,i}^{\text{mutvirus}} \mid N_r^{\text{mutvirus}}, \epsilon_{r,i}, \rho_{r,i}, \mu_{r,i}, \phi_{r,i}\right) = f\left(n_{r,i}^{\text{mutvirus}}; N_r^{\text{mutvirus}} \times [\epsilon_{r,i} + \rho_{r,i} + \mu_{r,i} \times \phi_{r,i}]\right) \tag{5}$$

We also specify priors over $\epsilon_{r,i}$, $\rho_{r,i}$, $\mu_{r,i}$, and $\phi_{r,i}$ in the form of gamma distributions. Specifically, let

$$g\left(x;\alpha,\beta\right) = \beta^{\alpha} \frac{1}{\Gamma\left(\alpha\right)} x^{\alpha-1} \exp\left(-x\beta\right) \tag{6}$$

denote the gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, where Γ is the gamma function. Note that the mean is given by

$$\overline{x} = \int_{x=0}^{\infty} x \times g\left(x;\alpha,\beta\right) dx = \alpha/\beta. \tag{7}$$

For all priors, we use a shape parameter of $\alpha = 4$ to give a moderately broad distribution.

For the prior over ϕ, we choose β such that the mean of the prior distribution corresponds to $\overline{\phi} = 0.1$, so that

$$\Pr\left(\phi_{r,i}\right) = g\left(\phi_{r,i};\alpha,\alpha/\overline{\phi}\right). \tag{8}$$

This choice of $\overline{\phi}$ is guided by the idea that we expect that most mutations will be deleterious and so have enrichment values substantially less than one.

For the priors over the mutation and error rates, we choose the rate parameter β such that the mean of the prior distribution is equal to the average value for the whole library (for $\mu_{r,i}$) or the average of all codon mutations in the library with that many nucleotide mutations (for $\epsilon_{r,i}$ and $\rho_{r,i}$). For example, if the average fraction of mutated codons in the **mutDNA** library minus the background from the **DNA** library is 6.1×10^3, and there are 63 mutant codons at each site, $\overline{\mu} = 6.1 \times 10^{-3}/63 = 9.7 \times 10^{-5}$, so for this library replicate we set the rate parameter to $\alpha/\overline{\mu}$. So for library replicate #1, we use a prior of

$$\Pr\left(\mu_{r,i}\right) = g\left(\mu_{r,i};\alpha,\alpha/\overline{\mu}\right). \tag{9}$$

For $\mu_{r,i}$ and $\rho_{r,i}$ we choose a different prior depending on the number of nucleotide changes in the codon mutation, since sequencing, PCR, and reverse-transcription errors are far more likely to lead to single-nucleotide codon changes than multiple-nucleotide codon changes. Specifically, let $\mathcal{M}\left(r,i\right)$ be the number of nucleotide changes in the mutation of site r from its wildtype identity to some non-wildtype codon i. For example, if the wildtype codon at position r is `GCA` then $\mathcal{M}\left(r,\texttt{ACA}\right) = 1$ and $\mathcal{M}\left(r,\texttt{ATA}\right) = 2$. If the error rate in (**DNA**) library is $\overline{\epsilon_1} = 5.8 \times 10^{-4}/9 = 6.4 \times 10^{-5}$ for single-nucleotide codon mutations, $\overline{\epsilon_2} = 8.7 \times 10^{-6}/27 = 3.2 \times 10^{-7}$ for two-nucleotide codon mutations, and $\overline{\epsilon_3} = 4.0 \times 10^{-6}/27 = 1.5 \times 10^{-7}$ for three-nucleotide codon mutations. So we use a prior of

$$\Pr\left(\epsilon_{r,i}\right) = g\left(\epsilon_{r,i};\alpha,\alpha/\overline{\epsilon_{\mathcal{M}(r,i)}}\right). \tag{10}$$

Similarly, the values for the reverse-transcription mutation rate for library replicate #1 (estimated from the **RNA** library minus the **DNA** library) are $\overline{\rho_1} = 1.9 \times 10^{-4}/9 = 2.1 \times 10^{-5}$, $\overline{\rho_2} = 1.5 \times 10^{-5}/27 = 5.6 \times 10^{-7}$, and $\overline{\rho_3} = 2.8 \times 10^{-6}/27 = 1.0 \times 10^{-7}$, and so we use a prior of

$$\Pr\left(\rho_{r,i}\right) = g\left(\rho_{r,i}; \alpha, \alpha/\overline{\rho_{\mathcal{M}(r,i)}}\right). \tag{11}$$

Given all of these likelihoods and priors, the overall posterior probability of a specific parameterization for the enrichment ratio and the unknown rates is given by

$$\begin{aligned}\Pr\left(\phi_{r,i}, \epsilon_{r,i}, \rho_{r,i}, \mu_{r,i} \mid \mathcal{N}_{r,i}\right) &= \mathcal{C}_{r,i} \times \Pr\left(n_{r,i}^{\text{DNA}} \mid N_r^{\text{DNA}}, \epsilon_{r,i}\right) \times \Pr\left(n_{r,i}^{\text{RNA}} \mid N_r^{\text{RNA}}, \epsilon_{r,i}, \rho_{r,i}\right) \times \\ &\quad \Pr\left(n_{r,i}^{\text{mutDNA}} \mid N_r^{\text{mutDNA}}, \epsilon_{r,i}, \mu_{r,i}\right) \times \\ &\quad \Pr\left(n_{r,i}^{\text{mutvirus}} \mid N_r^{\text{mutvirus}}, \epsilon_{r,i}, \rho_{r,i}, \mu_{r,i}, \phi_{r,i}\right) \times \\ &\quad \Pr\left(\epsilon_{r,i}\right) \times \Pr\left(\rho_{r,i}\right) \times \Pr\left(\mu_{r,i}\right) \times \Pr\left(\phi_{r,i}\right)\end{aligned} \tag{12}$$

where $\mathcal{C}_{r,i}$ is a normalization constant that does not need to be explicitly calculated in the approach used here, and $\mathcal{N}_{r,i} = \left\{n_{r,i}^{\text{DNA}}, n_{r,i}^{\text{mutDNA}}, n_{r,i}^{\text{RNA}}, n_{r,i}^{\text{mutvirus}}, N_r^{\text{DNA}}, N_r^{\text{mutDNA}}, N_r^{\text{RNA}}, N_r^{\text{mutvirus}}\right\}$ denotes the full set of counts for mutant codon i at site r.

We examine selection operating at the level of amino-acid rather than codon sequence, and so assume that the true value of enrichment ratio $\phi_{r,i}$ is equal for all codons i that encode the same amino acid at position r (this assumption is probably not completely accurate, and the study of differential enrichment for synonymous codons at a given site is an interesting area for future work). Let $\mathcal{A}_a$ denote the set of all codons for amino-acid a, and let $\phi_{r,a}$ denote the enrichment ratio for each codon encoding amino acid a (there is just one such enrichment ratio for all of these codons since we are assuming $\phi_{r,a} = \phi_{r,i}$ for all $i \in \mathcal{A}_a$). Then $\phi_{r,a}$ can be calculated from posterior probabilities defined in Equation 12 as

$$\Pr\left(\phi_{r,a} \mid \{\mathcal{N}_{r,i} \mid i \in \mathcal{A}_a\}\right) = \frac{\mathcal{C}_{r,a}}{\left[\Pr\left(\phi_{r,a}\right)\right]^{|\mathcal{A}|-1}} \times \prod_{i \in \mathcal{A}_a} \int_0^\infty \int_0^\infty \int_0^\infty \Pr\left(\phi_{r,a}, \epsilon_{r,i}, \rho_{r,i}, \mu_{r,i} \mid \mathcal{N}_{r,i}\right) d\epsilon_{r,i} d\rho_{r,i} d\mu_{r,i} \tag{13}$$

where $\mathcal{C}_{r,a}$ is again a normalization constant that does not need to be explicitly calculated in the approach used here, and where the $\left[\Pr\left(\phi_{r,a}\right)\right]^{|\mathcal{A}|-1}$ term ensures that the prior over $\phi_{r,a}$ is only included once in the calculation.

In practice, we compute the posterior distribution defined in Equation 13 using Markov Chain Monte Carlo (MCMC) over all of the unknown parameters ($\phi_{r,a}$ and all of the $\epsilon_{r,i}$, $\rho_{r,i}$, and $\mu_{r,i}$ values). We summarize the posterior distribution by its mean,

$$\langle \phi_{r,a} \rangle = \int_0^\infty \phi_{r,a} \times \Pr\left(\phi_{r,a} \mid \{\mathcal{N}_{r,i} \mid i \in \mathcal{A}_a\}_j\right) d\phi_{r,a}. \tag{14}$$

In our experiments, we perform several replicates of the experiment, and calculate a value for $\langle \phi_{r,a} \rangle$ for each of these replicates. For our final inferred values, we would like to summarize the inferred enrichments for all of the replicates. In principle, this could be done by extending Equation 14 to integrate over the posterior for several replicates. However, we instead prefer to summarize the enrichment by the geometric mean of the $\langle \phi_{r,a} \rangle$ values for the different libraries, since this approach is more robust to avoiding inflation of values due to an outlier with a large number of counts due to a source of error not included in the inference approach (such as linkage between mutations). The overall inferred enrichment ratio for $\mathcal{R}$ library replicates is then defined as

$$\overline{\langle \phi_{r,a} \rangle} = \left(\prod_{j=1}^{\mathcal{R}} \langle \phi_{r,a} \rangle_j\right)^{1/\mathcal{R}} \tag{15}$$

where $\langle \phi_{r,a} \rangle_j$ is the inferred enrichment ratio (Equation 14) for library replicate j.

We also calculate an equilibrium preference $\pi_{r,a}$ for each amino acid a (including the wildtype one) at site r as

$$\pi_{r,a} = \frac{\overline{\langle \phi_{r,a} \rangle}}{\sum_a \overline{\langle \phi_{r,a} \rangle}} \tag{16}$$

where we define $\overline{\langle \phi_{r,a} \rangle}$ to be one when a is the wildtype amino-acid at site r and by Equation 15 otherwise, and the summation is taken over all amino-acids a.

Finally, we calculate an estimated "entropy" for site r in bits as

$$h_r = -\sum_a \pi_{r,a} \times \log_2 \left(\pi_{r,a} \right) \tag{17}$$

where the sum is again taken over all amino-acids a.