Algorithm for enrichment of equilibrium preferences

Jesse D. Bloom[1*]

[1]Division of Basic Sciences and Computational Biology Program,
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[*]To whom correspondence should be addressed; E-mail: jbloom@fhcrc.org.

Quantifying the preference for amino acids at different sites

Conceptually, our approach is based on the idea that there is an inherent preference for each amino acid at each site in the protein. For now, we assume that these preferences are entirely at the amino-acid level and are indifferent to the specific codon (this assumption is probably not completely accurate, and the study of differential preferences for synonymous codons is an interesting area for future work). We denote the preference of site r for amino-acid a as $\pi_{r,a}$, where these equilibrium preferences are subject to the constraint

$$\sum_a \pi_{r,a} = 1, \tag{1}$$

and where a can range over the $n_{\rm aa} = 21$ values corresponding to all amino acids plus a stop codon (although in general we expect stop codons to be highly disfavored within the protein). We define the ratio $\pi_{r,a}/\pi_{r,a'}$ to be the expected ratio of amino-acid a to amino-acid a' after viral growth if both are initially introduced into the mutant library at equal frequency. Mutations that enhance viral growth will have larger values of $\pi_{r,a}$, while mutations that hamper viral growth will have lower values of $\pi_{r,a}$. However, note that $\pi_{r,a}/\pi_{r,a'}$ cannot be simply interpreted as the fitness effect of mutating site r from a to a': because most clones in our mutant libraries have multiple mutations, this ratio summarizes the effect of a mutation in the wildtype gene and a variety of closely related mutants. A mutation can therefore have a ratio greater than one due to its inherent effect on viral growth or its effect on the tolerance of the protein for other mutations (or a combination of both). The analysis here does not separate these two factors, but note that previous experimental work has shown that it is fairly common for one mutation in NP to alter the protein's tolerance to a subsequent mutation.

The most naive approach is to set $\pi_{r,a}$ proportional to the frequency of amino-acid a in the **mutvirus-p1** library divided by the frequency of the mutation in the **mutDNA** library, and then

apply the normalization condition in Equation 1 to get the proportionality constant. However, such an approach is problematic for several reasons. First, it fails to account for other sources of error and mutation (PCR, reverse-transcription, etc) that inflate the observed frequencies of some mutations. Second, the libraries contain finite numbers of counts, 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-p1** library and not at all in the **mutDNA** library, the ratio is infinity – yet in practice such low counts give us little confidence that the enough variants have been assayed to estimate the true effect of the mutation.

To circumvent these problems, we use a Bayesian approach that explicitly accounts for the sampling statistics. We begin with prior estimates that the error and mutation rates for each site are equal to the library averages. We specify likelihood functions that give the probability of observing a set of counts given the $\pi_{r,a}$ values for that site and the various error and mutation rates. We then estimate the posterior distribution of the $\pi_{r,a}$ values by MCMC. This approach accounts for sources of error and avoids overfitting $\pi_{r,a}$ when the number counts is low.

We use the counts in the **DNA** library to quantify errors due to PCR 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. The second of these assumptions is supported by the fact that the total mutation frequency in the **virus-p1** libraries is close to that in the **RNA** libraries. The first of these assumptions is supported by the fact that stop codons are no more frequent in the **RNA** libraries than in the **virus-p1** libraries – deleterious stop codons arising during transcription will be purged during viral growth, while those arising from reverse-transcription and sequencing errors will not.

At each codon site r, there are $n_{\text{codon}} = 64$ codons, which we index by $i = 1, 2, \ldots n_{\text{codon}}$. Let $\text{wt}\,(r)$ denote the wildtype codon at site r. Let N_r^{DNA} be the total number of sequencing reads at site r in the **DNA** library, and let $n_{r,i}^{\text{DNA}}$ be the number of these reads that report codon i at site r, so that $\sum_i n_{r,i}^{\text{DNA}} = N_r^{\text{DNA}}$. Similarly, let N_r^{mutDNA}, N_r^{RNA}, and N_r^{mutvirus} be the total number of reads at site r and let $n_{r,i}^{\text{mutDNA}}$, $n_{r,i}^{\text{RNA}}$, and $n_{r,i}^{\text{mutvirus}}$ be the total number of these reads that report codon i at site r in the **mutDNA**, **RNA**, and **mutvirus-p1**, respectively.

We first consider the rate at which site r is erroneously read to be some incorrect identity due to PCR or sequencing errors. Such errors are the only source of non-wildtype reads in the sequencing of the **DNA** library. For all $i \neq \text{wt}\,(r)$, we define $\epsilon_{r,i}$ as the rate at which site r is erroneously read as codon i in the **DNA** library. We define $\epsilon_{r,\text{wt}(r)} = 1 - \sum_{i \neq \text{wt}(r)} \epsilon_{r,i}$ to be the rate at which site r is correctly read as its wildtype identity of $\text{wt}\,(r)$ in the **DNA** library. We therefore have $\epsilon_{r,i} = \mathbb{E}\left[n_{r,i}^{\text{DNA}}/N_r^{\text{DNA}}\right]$ where $\mathbb{E}$ denotes the expectation value. If we define $\overrightarrow{\epsilon_r} = (\epsilon_{r,1}, \ldots, \epsilon_{r,n_{\text{codon}}})$ and $\overrightarrow{n_r^{\text{DNA}}} = \left(n_{r,1}^{\text{DNA}}, \ldots, n_{r,n_{\text{codon}}}^{\text{DNA}}\right)$ as vectors of the $\epsilon_{r,i}$ and $n_{r,i}^{\text{DNA}}$ values, then the likelihood of observing $\overrightarrow{n_r^{\text{DNA}}}$ given $\overrightarrow{\epsilon_r}$ and N_r^{DNA} is

$$\Pr\left(\overrightarrow{n_r^{\text{DNA}}} \mid N_r^{\text{DNA}}, \overrightarrow{\epsilon_r}\right) = \text{Mult}\left(\overrightarrow{n_r^{\text{DNA}}}; N_r^{\text{DNA}}, \overrightarrow{\epsilon_r}\right) \tag{2}$$

where Mult denotes the multinomial distribution.

We next consider the rate at which site r is erroneously copied during reverse transcription. These reverse-transcription errors combine with the PCR / sequencing errors defined by $\overrightarrow{\epsilon_r}$ to create non-wildtype reads in the **RNA** library. For all $i \neq \mathrm{wt}\,(r)$, we define $\rho_{r,i}$ as the rate at which site r is miscopied to codon i during reverse transcription. We define $\rho_{r,\mathrm{wt}(r)} = 1 - \sum_{i \neq \mathrm{wt}(r)} \rho_{r,i}$ as the rate at which site r is correctly reverse transcribed. If we ignore as negligibly rare the possibility that a site is subject to both a reverse-transcription and sequencing / PCR error within the same clone (a reasonable assumption as both $\epsilon_{r,i}$ and $\rho_{r,i}$ are very small for $i \neq \mathrm{wt}\,(r)$), then $\epsilon_{r,i} + \rho_{r,i} - \delta_{i,\mathrm{wt}(r)} = \mathbb{E}\left[n_{r,i}^{\mathrm{RNA}} / N_r^{\mathrm{RNA}}\right]$ where $\delta_{i,\mathrm{wt}(r)}$ is the Kronecker delta (equal to one if $i = \mathrm{wt}\,(r)$ and zero otherwise). The likelihood of observing $\overrightarrow{n_r^{\mathrm{RNA}}}$ given $\overrightarrow{\rho_r}$, $\overrightarrow{\epsilon_r}$, and N_r^{RNA} is

$$\Pr\left(\overrightarrow{n_r^{\mathrm{RNA}}} \mid N_r^{\mathrm{RNA}}, \overrightarrow{\rho_r}, \overrightarrow{\epsilon_r}\right) = \mathrm{Mult}\left(\overrightarrow{n_r^{\mathrm{RNA}}}; N_r^{\mathrm{RNA}}, \overrightarrow{\epsilon_r} + \overrightarrow{\rho_r} - \overrightarrow{\delta_r}\right). \tag{3}$$

where $\overrightarrow{\delta_r} = \left(\delta_{1,\mathrm{wt}(r)}, \ldots, \delta_{n_{\mathrm{codon}},\mathrm{wt}(r)}\right)$ is a vector that is all zeros except for the element corresponding to $\mathrm{wt}\,(r)$.

We next consider the rate at which site r is mutated to some other codon in the plasmid mutant library. These mutations combine with the PCR / sequencing errors defined by $\overrightarrow{\epsilon_r}$ to create non-wildtype reads in the **mutDNA** library. For all $i \neq \mathrm{wt}\,(r)$, we define $\mu_{r,i}$ as the rate at which site r is mutated to codon i in the mutant library. We define $\mu_{r,\mathrm{wt}(r)} = 1 - \sum_{i \neq \mathrm{wt}(r)} \mu_{r,i}$ as the rate at which site r is not mutated. If we ignore as negligibly rare the possibility that a site is subject to both a mutation and a sequencing / PCR error within the same clone, then $\mu_{r,i} + \epsilon_{r,i} - \delta_{i,\mathrm{wt}(r)} = \mathbb{E}\left[n_{r,i}^{\mathrm{mutDNA}} / N_r^{\mathrm{mutDNA}}\right]$. The likelihood of observing $\overrightarrow{n_r^{\mathrm{mutDNA}}}$ given $\overrightarrow{\mu_r}$, $\overrightarrow{\epsilon_r}$, and N_r^{mutDNA} is

$$\Pr\left(\overrightarrow{n_r^{\mathrm{mutDNA}}} \mid N_r^{\mathrm{mutDNA}}, \overrightarrow{\mu_r}, \overrightarrow{\epsilon_r}\right) = \mathrm{Mult}\left(\overrightarrow{n_r^{\mathrm{mutDNA}}}; N_r^{\mathrm{mutDNA}}, \overrightarrow{\mu_r} + \overrightarrow{\epsilon_r} - \overrightarrow{\delta_r}\right). \tag{4}$$

Finally, we consider the effect of the preferences of each site r for different amino acids, as denoted by the $\pi_{r,a}$ values. Selection due to these preferences is manifested in the **mutvirus** library. This selection acts on the mutations in the mutant library ($\mu_{r,i}$), although the actual counts in the **mutvirus** library are also affected by the sequencing / PCR errors ($\epsilon_{r,i}$) and reverse-transcription errors ($\rho_{r,i}$). We again ignore as negligibly rare the possibility that a site is subject to more than one of these sources of mutation and error within a single clone. Let $\mathcal{A}\,(i)$ denote the amino acid encoded by codon i. Let $\overrightarrow{\pi_r}$ be the vector of $\pi_{r,a}$ values. Define the vector-valued function $\overrightarrow{\mathcal{C}}$ as

$$\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right) = \left(\pi_{r,\mathcal{A}(1)}, \ldots, \pi_{r,\mathcal{A}(n_{\mathrm{codon}})}\right), \tag{5}$$

so that this function returns a n_{codon}-element vector constructed from $\overrightarrow{\pi_r}$. Because the selection in the **mutvirus** library due to the preferences $\pi_{r,\mathcal{A}(i)}$ occurs after the mutagenesis $\mu_{r,i}$ but before the reverse-transcription errors $\rho_{r,i}$ and the sequencing / PCR errors $\epsilon_{r,i}$, we have

$\mathbb{E}\left[n_{r,i}^{\text{mutvirus}}/N_r^{\text{mutvirus}}\right] = \epsilon_{r,i}+\rho_{r,i}+\gamma_r\times\pi_{r,\mathcal{A}(i)}\times\mu_{r,i}-2\times\delta_{i,\text{wt}(r)}$ where $\gamma_r = \left(\sum_i \pi_{r,\mathcal{A}(i)}\mu_{r,i}\right)^{-1} = \left(\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right)\cdot\overrightarrow{\mu_r}\right)^{-1}$ (where $\cdot$ denotes the dot product) is a normalization factor that accounts for the fact that changes in the frequency of one variant due to selection will influence the observed frequency of other variants. The likelihood of observing $\overrightarrow{n_r^{\text{mutvirus}}}$ given $\overrightarrow{\mu_r}$, $\overrightarrow{\epsilon_r}$, $\overrightarrow{\rho_r}$, $\overrightarrow{\pi_r}$, and N_r^{mutvirus} is therefore

$$\Pr\left(\overrightarrow{n_r^{\text{mutvirus}}} \mid \overrightarrow{\mu_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\pi_r}, N_r^{\text{mutvirus}}\right) = \text{Mult}\left(\overrightarrow{n_r^{\text{mutvirus}}}; N_r^{\text{mutvirus}}, \overrightarrow{\epsilon_r} + \overrightarrow{\rho_r} + \frac{\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right)\circ\overrightarrow{\mu_r}}{\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right)\cdot\overrightarrow{\mu_r}} - 2\overrightarrow{\delta_r}\right). \tag{6}$$

where $\circ$ is the Hademard (entry-wise) product.

We specify priors over $\overrightarrow{\epsilon_r}$, $\overrightarrow{\rho_r}$, $\overrightarrow{\mu_r}$, and $\overrightarrow{\pi_r}$ in the form of Dirichlet distributions (denoted here by Dir). For the priors over the mutation rates $\overrightarrow{\mu_r}$, we choose Dirichlet-distribution parameters such that the mean of the prior expectation for the mutation rate at each site r and codon i is proportional to the average value for all sites. The average fraction of mutated codons in the **mutDNA** library minus the background from the **DNA** library is 6.1×10^{-3}, so the average mutation rate is $\overline{\mu} = 6.1 \times 10^{-3}/63 = 9.7 \times 10^{-5}$. So we use a prior of

$$\Pr\left(\overrightarrow{\mu_r}\right) = \text{Dir}\left(\overrightarrow{\mu_r}; n_{\text{codon}} \cdot \sigma_\mu \cdot \overrightarrow{\alpha_{\mu,r}}\right) \tag{7}$$

where $\overrightarrow{\alpha_{\mu,r}}$ is the n_{codon}-element vector with elements $\alpha_{\mu,r,i} = \overline{\mu} + \delta_{i,\text{wt}(r)}\left(1 - n_{\text{codon}}\overline{\mu}\right)$ and σ_μ is the scalar concentration parameter. For a symmetric Dirichlet distribution (which $\Pr\left(\overrightarrow{\mu_r}\right)$ will in general *not* be), choosing $\sigma_\mu = 1$ makes the distribution completely uniform.

For the priors over $\epsilon_{r,i}$ and $\rho_{r,i}$, the Dirichlet-distribution parameters again represent the average value for all sites, but now also depend 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. Let $\mathcal{M}\left(\text{wt}\left(r\right), i\right)$ be the number of nucleotide changes in the mutation from codon $\text{wt}\left(r\right)$ to codon i. For example, $\mathcal{M}\left(\texttt{GCA}, \texttt{ACA}\right) = 1$ and $\mathcal{M}\left(\texttt{GCA}, \texttt{ATA}\right) = 2$. The average error rate (estimated from the **DNA** library) is $\overline{\epsilon_1} = 5.8 \times 10^{-4}/9 = 6.4 \times 10^{-5}$ for single-nucleotide codon mutations, $\overline{\epsilon_2} = 8.7\times10^{-6}/27 = 3.2\times10^{-7}$ for two-nucleotide codon mutations, and $\overline{\epsilon_3} = 4.0\times10^{-6}/27 = 1.5 \times 10^{-7}$ for three-nucleotide codon mutations. So we use a prior of

$$\Pr\left(\overrightarrow{\epsilon_r}\right) = \text{Dir}\left(\overrightarrow{\epsilon_r}; n_{\text{codon}} \cdot \sigma_\epsilon \cdot \overrightarrow{\alpha_{\epsilon,r}}\right) \tag{8}$$

where $\overrightarrow{\alpha_{\epsilon,r}}$ is the n_{codon}-element vector with elements $\alpha_{\epsilon,r,i} = \overline{\epsilon_{\mathcal{M}(\text{wt}(r),i)}}$ where we define $\overline{\epsilon_0} = 1 - 9 \times \overline{\epsilon_1} - 27 \times \overline{\epsilon_2} - 27 \times \overline{\epsilon_3}$ (for any codon, there are 9 one-nucleotide mutations, 27 two-nucleotide mutations, and 27 three-nucleotide mutations), and where σ_ϵ is the scalar concentration parameter.

Similarly, the average reverse-transcription error rates (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 $\overline{\rho_3} = 2.8 \times 10^{-6}/27 = 1.0 \times 10^{-7}$ for one-, two-, and three-nucleotide codon mutations, respectively. So we use a prior of

$$\Pr\left(\overrightarrow{\rho_r}\right) = \mathrm{Dir}\left(\overrightarrow{\rho_r}; n_{\mathrm{codon}} \cdot \sigma_\rho \cdot \overrightarrow{\alpha_{\rho,r}}\right) \tag{9}$$

where $\overrightarrow{\alpha_{\rho,r}}$ is the n_{codon}-element vector with elements $\alpha_{\rho,r,i} = \overline{\rho_{\mathcal{M}(\mathrm{wt}(r),i)}}$ where we define $\overline{\rho_0} = 1 - 9 \times \overline{\rho_1} - 27 \times \overline{\rho_2} - 27 \times \overline{\rho_3}$, and where σ_ρ is the scalar concentration parameter.

We specify a symmetric Dirichlet-distribution prior over $\overrightarrow{\pi_r}$ (note that any other prior, such as one that favored the wildtype identity, would implicitly favor certain identities based empirically on the wildtype protein sequence, and so would not be in the spirit of the parameter-free derivation of the $\pi_{r,a}$ values employed here). Specifically, we use a prior of

$$\Pr\left(\overrightarrow{\pi_r}\right) = \mathrm{Dir}\left(\overrightarrow{\pi_r}; \sigma_\pi \cdot \overrightarrow{1}\right) \tag{10}$$

where $\overrightarrow{1}$ is the n_{aa}-element vector that is all ones, and σ_π is the scalar concentration parameter. Setting $\sigma_\pi = 1$ gives a uniform prior over all $\overrightarrow{\pi_r}$ values, while smaller values favor peaked distribution and larger values favor equal values for all $\pi_{r,a}$ elements.

We can now write expressions for the likelihoods and posterior probabilities. Let $\mathcal{N}_r = \left\{\overrightarrow{n_r^{\mathrm{DNA}}}, \overrightarrow{n_r^{\mathrm{mutDNA}}}, \overrightarrow{n_r^{\mathrm{RNA}}}, \overrightarrow{n_r^{\mathrm{mutvirus}}}, N_r^{\mathrm{DNA}}, N_r^{\mathrm{mutDNA}}, N_r^{\mathrm{RNA}}, N_r^{\mathrm{mutvirus}}\right\}$ denote the full set of counts for site r. The likelihood of $\mathcal{N}_r$ given values for the equilibrium preferences and mutation / error rates is

$$\begin{aligned}\Pr\left(\mathcal{N}_r \mid \overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r}\right) &= \Pr\left(\overrightarrow{n_r^{\mathrm{DNA}}} \mid N_r^{\mathrm{DNA}}, \overrightarrow{\epsilon_r}\right) \times \Pr\left(\overrightarrow{n_r^{\mathrm{RNA}}} \mid N_r^{\mathrm{RNA}}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}\right) \times \\ &\quad \Pr\left(\overrightarrow{n_r^{\mathrm{mutDNA}}} \mid N_r^{\mathrm{mutDNA}}, \overrightarrow{\epsilon_r}, \overrightarrow{\mu_r}\right) \times \\ &\quad \Pr\left(\overrightarrow{n_r^{\mathrm{mutvirus}}} \mid N_r^{\mathrm{mutvirus}}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r}, \overrightarrow{\pi_r}\right) \end{aligned} \tag{11}$$

where the likelihoods that compose Equation 11 are defined by Equations 2, 3, 4, and 6. The posterior probability of a specific value for the equilibrium preferences and mutation / error rates is

$$\begin{aligned}\Pr\left(\overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r} \mid \mathcal{N}_r\right) &= C_r \times \Pr\left(\mathcal{N}_r \mid \overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r}\right) \times & (12) \\ &\quad \Pr\left(\overrightarrow{\epsilon_r}\right) \times \Pr\left(\overrightarrow{\rho_r}\right) \times \Pr\left(\overrightarrow{\mu_r}\right) \times \Pr\left(\overrightarrow{\pi_r}\right) & (13)\end{aligned}$$

where C_r is a normalization constant that does not need to be explicitly calculated in the MCMC approach used here. The posterior over the equilibrium preferences $\overrightarrow{\pi_r}$ can be calculated by integrating over Equation 12 to give

$$\Pr\left(\overrightarrow{\pi_r} \mid \mathcal{N}_r\right) = \int\int\int \Pr\left(\overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r} \mid \mathcal{N}_r\right) d\overrightarrow{\epsilon_r} d\overrightarrow{\rho_r} d\overrightarrow{\mu_r}, \tag{14}$$

where the integration is performed by MCMC.

Equation 14 infers $\overrightarrow{\pi_r}$ for a single replicate of the experiment. In practice, we have performed $\mathcal{R} = 4$ replicates (WT-1, WT-2, N334H-1, N334H-2). Let $\mathcal{N}_r^k$ denote the set of experimentally observed counts for replicate k. The full set of data for all $\mathcal{R}$ replicates is then $\left\{\mathcal{N}_r^k \mid 1 \leq k \leq \mathcal{R}\right\}$. We use the same equilibrium preferences $\overrightarrow{\pi_r}$ for all replicates, since these preferences are a property of the protein rather than the experiment. However, the error and mutation rates $\overrightarrow{\epsilon_r}$, $\overrightarrow{\mu_r}$, and $\overrightarrow{\rho_r}$ are specific to each replicate (denote the replicate-specific vectors by $\overrightarrow{\epsilon_r}^k$, $\overrightarrow{\mu_r}^k$, and $\overrightarrow{\rho_r}^k$), with the replicate-specific prior vectors $\overrightarrow{\alpha_{\mu,r}}^k$, $\overrightarrow{\alpha_{\epsilon,r}}^k$, and $\overrightarrow{\alpha_{\rho,r}}^k$ determined from the library averages for that replicate. The reason for making these three vectors and their priors replicate specific is that each replicate could have different error and mutation rates. The posterior probability of $\overrightarrow{\pi_r}$ and the mutation / error rates for all experimental replicates is

$$\begin{aligned}\Pr\left(\overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r} \mid \left\{\mathcal{N}_r^k \mid 1 \leq k \leq \mathcal{R}\right\}\right) &= C_r' \times \prod_{k=1}^{\mathcal{R}} \Big[\Pr\left(\mathcal{N}_r^k \mid \overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}^k, \overrightarrow{\rho_r}^k, \overrightarrow{\mu_r}^k\right) \times \\ &\quad \Pr\left(\overrightarrow{\epsilon_r}^k\right) \Pr\left(\overrightarrow{\rho_r}^k\right) \Pr\left(\overrightarrow{\mu_r}^k\right)\Big] \Pr\left(\overrightarrow{\pi_r}\right), \quad (15)\end{aligned}$$

where each of the individual likelihoods for the $\mathcal{N}_r^k$ counts are calculated using Equation 11 and C_r' is a normalization constant that does not need to be calculated when using MCMC. The posterior over the equilibrium preferences $\overrightarrow{\pi_r}$ can be calculated by using MCMC to integrate over Equation 15 to give

$$\Pr\left(\overrightarrow{\pi_r} \mid \left\{\mathcal{N}_r^k \mid 1 \leq k \leq \mathcal{R}\right\}\right) = \int \cdots \int \Pr\left(\overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r} \mid \left\{\mathcal{N}_r^k \mid 1 \leq k \leq \mathcal{R}\right\}\right) \prod_{k=1}^{\mathcal{R}} d\overrightarrow{\epsilon_r}^k d\overrightarrow{\rho_r}^k d\overrightarrow{\mu_r}^k. \quad (16)$$

We summarize the posterior calculated from Equation 16 by its mean,

$$\langle \overrightarrow{\pi_r} \rangle = \int \overrightarrow{\pi_r} \times \Pr\left(\overrightarrow{\pi_r} \mid \left\{\mathcal{N}_r^k \mid 1 \leq k \leq \mathcal{R}\right\}\right) d\overrightarrow{\pi_r}. \quad (17)$$