# Estimating Selection and Mutation Rates from a Random Field Model for Polymorphic Sites

Stanley A. Sawyer *,†

Selection and mutation rates are estimated from aligned DNA sequences by using a Poisson random field model for base frequencies at polymorphic sites. This approach is applied at amino acid polymorphic codon positions to estimate selection rates against observed amino acid polymorphisms and also the proportion of codon positions that admit a weakly selected replacement. The technique can also be used at silent sites to obtain a numerical estimate of codon bias, which is shown in two examples to be of comparable strength to the selection against observed variant amino acids. A nonparametric test that can detect selection against amino acids using silent sites as a control is also discussed. Finally, a technique that estimates silent mutation rates assuming selective neutrality of silent polymorphisms, but which is not sensitive to saturation, is used as a consistency check.

## 1. Introduction

The basic genetic material or DNA of plants and animals is composed of one or more chromosomes, where the actual number of chromosomes depends on the plant or animal. Each chromosome can be thought of as a long string of letters from the alphabet $\mathrm{T}, \mathrm{C}, \mathrm{A}, \mathrm{G}$, where each letter corresponds to a specific nucleotide. In this sense, a mouse is the same as a tomato to a geneticist, since both have about the same amount of DNA.[1] A gene or genetic locus is a segment of a chromosome that affects some trait. Typically, a gene contains one or more coding regions for a protein or RNA type along with some recognition sites for regulatory molecules. Some of the most important proteins are enzymes, which are proteins that catalyze biochemical reactions.

Beginning in the 1960's, many experiments found that enzymes in human and natural populations were often polymorphic; i.e., had multiple forms in the

* Washington University, St. Louis, MO 63130, USA

† This work was partially supported by National Science Foundation Grant DMS-9108262 and National Institutes of Health research grant GM-44889.

[1] Attributed to Eric Lander.

population (HARRIS, 1966; LEWONTIN and HUBBY, 1966). Some biologists felt that most of this variation is selectively neutral, with at most negligible effects on fitness (KIMURA, 1983). That is, whichever form of the enzyme that a creature possessed made no difference to its survival or lifestyle. Other biologists felt that different enzymes were unlikely to be selectively equivalent, and so various forms of selection must be involved (WILLS, 1973; LEWONTIN, 1974; *see also* HARTL, 1989; HARTL and SAWYER, 1991).

The coding regions of genes in DNA are directly translated into strings of amino acids by various RNA enzymes. Proteins and protein enzymes are built up from one or more strings of amino acids. Triples of nucleotides from a coding region of a gene are called *codons*. There are $4^3 = 64$ possible codons, of which 61 code for the 20 different amino acids and the remaining 3 indicate the end of a string of amino acids. Most amino acids are coded for by either two or four distinct codons, with the variability in the third nucleotide position. Thus the two codons CAT and CAC (and no other codons) code for the amino acid histine, and the four codons GTT, GTC, GTA, and GTG code for valine. A change in the DNA in a coding region that does not change the corresponding amino acid (for example, a change from CAT to CAC or from GTT to GTA) is called a *silent* or *synonymous* change. If a change in the DNA causes a change in the amino acid, it is called a *replacement* change. (The term "replacement" in this context refers only to amino acids.) As one might expect, DNA from natural populations shows a great deal of silent as well as replacement variation.

Curiously, the different codons for the same amino acid tend to occur in different frequencies (HARTL and CLARK, 1989; LI and GRAUR, 1991). This tendency is called *codon bias*. One possible reason for codon bias could be biased mutation rates at the DNA level. That is, some of the nucleotides $\mathrm{T}, \mathrm{C}, \mathrm{A}, \mathrm{G}$ that make up DNA may be more mutable than others, or else mutation may preferentially result in some nucleotides as opposed to others. On the other hand, selection may also be involved. Some codons may be necessary for the proper configuration of DNA. The transfer RNA's that are necessary for each codon to translate DNA occur in different frequencies for different codons for the same amino acid, which may slow the translation of some synonymous codons. However, rates of nucleotide substitution at silent codon sites are similar to those in pseudogenes (LI *et al.* 1985; WOLFE *et al.* 1989). (A pseudogene is a DNA segment that resembles a gene but is not expressed.) This suggests that most silent substitutions are selectively neutral or nearly selectively neutral. While in general silent DNA changes appear to be under weaker selective constraints than changes that cause amino acid replacement, some silent changes are known that have a selective effect.

The purpose here is to discuss methods for detecting and estimating se-

lection based on an aligned set of DNA sequences. The main technique that we discuss is a random field model for the frequencies of variant nucleotides at the various polymorphic sites that allows quantitative estimates of both selection and mutation rates (Section 4). This approach can be applied independently to silent polymorphisms and to amino acid variation, and provides independent estimates for both the selection against replacement substitutions and the selection involved in codon bias.

We begin with a nonparametric test for selection against replacement substitutions that uses silent substitutions as a selectively neutral "control" (Section 2). While this is the simplest of the three methods that we discuss, it does not give quantitative estimates of selection or mutation, and cannot detect selection against replacement substitutions if selection against silent substitutions is equally strong, which appears to be the case in one of our two examples. Section 3 is devoted to an independent method for estimating mutation rates at silent sites under different assumptions that provides a consistency check for the random field model.

A fringe benefit of the random field analysis is that you can estimate separate gene locus-wide mutation rates for silent changes and for amino-acid replacements. Many changes to a functioning protein or enzyme are presumably lethal or nearly lethal. Amino-acid variation that is common enough to be detected in a sample of DNA sequences must be subject to relatively weak selection. An estimate of the ratio of the amino-acid replacement mutation rate to the silent mutation rate should give an estimate of the proportion of amino acids in a protein that can be replaced without lethal effects on the host.

The models discussed below all assume that changes to an ancestral base or amino acid are either strongly deleterious (and so will never be seen in a natural population) or else change the fitness of the host by an equal amount, equal both for changes to different bases at the same site and for changes at different sites. Mutations at different sites have multiplicative effects on fitness. In particular, these models are not designed to detect "balancing" selection, in which rare enzymes have a selective advantage, or selection which is nonmultiplicative across sites. Of course, no statistical test can detect *arbitrary* forms of selection, since *anything* that you observe could be the result of selection for exactly that configuration.

## 2. A Contingency-Table Test for Unidirectional Selection

Suppose that you have $n$ aligned DNA sequences from a coding region. Most aligned sites will have a single most common base with at most a few sequences with different bases at that site. If most changes from this consensus base are selectively deleterious, then you would expect relatively few sequences at any

particular site to be different from the consensus base. Similarly, if changes from the consensus base were advantageous, then you would also expect relatively few differences from the consensus, since otherwise the consensus would be quickly driven from the population.

If polymorphic replacement sites are observed to be more bunched (i.e., have fewer deviations from the consensus) than polymorphic silent sites, then one possible explanation is unidirectional selection against amino acid replacements.

Specifically, given $n$ aligned DNA sequences, we say that a site is *simply polymorphic* if $n-1$ of the sequences have one base at this site and one sequence has a second base. All other polymorphic sites are called *multiply polymorphic*. A three-site codon is called *regular* if the first two nucleotide positions are nondegenerate (i.e., any replacement changes the amino acid). This is equivalent to saying that the codon corresponds to an amino acid other than leucine or arginine, which are the only two amino acids for which a single change in the first or second nucleotide position of a codon will not change the amino acid.

About half of regular amino acids are fourfold degenerate at the third position, which means that any base can be substituted at the third codon site without changing the amino acid. Most of the other amino acids are twofold degenerate at the third position. Two-fold degenerate amino acids are of two types. The first type can have either T or C at the third position, but any other change at the third site changes the amino acid. For the second type, the third base is either A or G. There is one threefold degenerate amino acid (isoleucine), which corresponds to the three codons ATT, ATC, and ATA. Since the codon ATA is extremely rare in most natural populations, we treat isoleucine as twofold degenerate and the rare codon positions with an ATA as irregular.

Now, consider a $2 \times 2$ contingency table with the numbers of silent simply and multiply polymorphic sites at amino-acid monomorphic regular codon positions in the first row, and the numbers of simply and multiply polymorphic sites at the first and second positions of regular codons in the second row (SAWYER, DYKHUIZEN, and HARTL, 1987; see Table 1). Regular codon positions have the potential of supplying two replacement polymorphisms, but this is rare, and historically may have been the result of two distinct amino-acid replacements.

The contingency table in Table 1 is highly significant for 14 strains of *Escherichia coli* at the *gnd* locus when all silent sites are used, but is not significant for 8 strains of *Salmonella typhimurium* at the *PutP* locus. Thus the selective forces affecting variant bases appear to be different for silent and replacement sites in *gnd*, but not in *PutP*.

**Table 1: 2 × 2 tables for selection**

14 strains of *Escherichia coli* at the *gnd* locus (1407bp): [a,b]

| | All silent[c] | | Two-fold silent[c] | |
|---|---|---|---|---|
| | simple poly | multiple poly | simple poly | multiple poly |
| Silent (regular)[c] | 60 | 83 | 27 | 31 |
| Replacements (1,2 pos'n regular) | 20 | 7 | 20 | 7 |
| | $P = 0.003$[f] | | $P = 0.021$[f] | |

8 strains of *Salmonella typhimurium* at the *PutP* locus (1467bp): [d,e]

| | | | | |
|---|---|---|---|---|
| Silent (regular) | 93 | 59 | 43 | 26 |
| Replacements (1,2 pos'n regular) | 12 | 4 | 12 | 4 |
| | $P = 0.416$[f] | | $P = 0.398$[f] | |

a – The *gnd* locus transcribes the enzyme 6-phosphogluconate dehydrogenase.

b – Dykhuizen and Green (1991); Bisercic, Feutrier, and Reeves (1991).

c – See text for definitions.

d – The *PutP* locus transcribes the enzyme proline permease.

e – Nelson and Selander (1992).

f – Two-sided Fisher exact test.

Most replacement variation may be lethal or at least subject to highly deleterious selection, so that there may be at most two weakly-selected bases at any amino-acid varying site. Thus it may be fairer to compare replacement polymorphisms with twofold degenerate silent polymorphisms, which are more likely to be simply polymorphic than fourfold degenerate sites. When twofold degenerate sites are used, the contingency table in Table 1 is significant for *gnd* but is not highly significant.

Note this test cannot detect selection against amino acid replacements if there is the same amount of selection against silent differences due to codon bias. Similarly, a positive result for this test might even be due to positive selection for silent nucleotide variants in combination with selectively neutral

amino acid variation.

## 3. Estimating Mutation Rates Assuming Silent Sites are Neutral

There are many different ways to estimate the amount of mutation at silent sites in an aligned set of DNA sequences (see e.g. FU and LI, 1993). The following approach has the advantage that it automatically allows for parallel or repeated mutations at the same site, and can also be adapted to estimate the divergence time between two species (SAWYER and HARTL, 1992; *see also* SAWYER, DYKHUIZEN, and HARTL, 1987).

Assume that a fourfold degenerate site (for example) has mutation rates $\mu_T, \mu_C, \mu_A$, and $\mu_G$ *to that base* per chromosome per generation. Thus the mutation rate depends on the base, but depends only on the end-product base (TAJIMA and NEI, 1982). Note that "mutations" of e.g. T to T that do not change the base are permitted here, but will not be counted below when estimating the locus-wide silent mutation rate.

Under these conditions, the *population* frequencies of the four bases at that site will have the joint probability density

$$C_A \, p_T^{\alpha_T - 1} \, p_C^{\alpha_C - 1} \, p_A^{\alpha_A - 1} \, p_G^{\alpha_G - 1} \, dp_T dp_C dp_A dp_G \tag{1}$$

where $\alpha_T = 2N_e\mu_T$, $\alpha_C = 2N_e\mu_C$, ..., where $N_e$ is the haploid effective population size and $C_A = C(\alpha_T, \alpha_C, \ldots)$, under the usual conditions for diffusion approximations (WRIGHT, 1949; KINGMAN, 1980). The density in equation (1) is called a Dirichlet density. The corresponding density for TC-twofold degenerate sites is the beta density $C'_A \, p_T^{\alpha_T - 1} \, p_C^{\alpha_C - 1} \, dp_T dp_C$ for $p_T + p_C = 1$, with a similar expression for AG-twofold degenerate sites.

Now assume that the site is part of an aligned sample of $n$ DNA sequences, and consider the probability that the sample has $n_T$ sequences with the base T at that site, $n_C$ sequences with C, $n_A$ with A, ..., where $n = n_T + n_C + n_A + n_G$. This probability can be obtained by integrating the density in equation (1), and is

$$C_N \, \frac{\alpha_T^{(n_T)} \alpha_C^{(n_C)} \alpha_A^{(n_A)} \alpha_G^{(n_G)}}{\alpha^{(n)}}, \qquad \alpha = \alpha_T + \alpha_C + \alpha_A + \alpha_G \tag{2}$$

where $x^{(k)} = x(x+1)\ldots(x+k-1)$ and $C_N = n!/(n_T!\, n_C!\, n_A!\, n_G!)$ (WATTERSON, 1977). The corresponding probability for TC-twofold degenerate sites is $C'_N \alpha_T^{(n_T)} \alpha_C^{(n_C)} / (\alpha_T + \alpha_C)^{(n)}$. The probabilities in equation (2) for fourfold degenerate sites, and the corresponding probabilities at the two different types

**Table 2: Maximum likelihood estimates of $\alpha_T, \alpha_C, \alpha_A, \alpha_G$ from the probabilities (2) at silent sites**

| *gnd*[a] | | ADH[b] | |
|---|---|---|---|
| alpha's | 4-fold[c] | alpha's | 4-fold[c] |
| $\alpha_T = 0.128$ | 0.407 | $\alpha_T = 0.0080$ | 0.155 |
| $\alpha_C = 0.109$ | 0.288 | $\alpha_C = 0.0300$ | 0.610 |
| $\alpha_A = 0.063$ | 0.106 | $\alpha_A = 0.0023$ | 0.066 |
| $\alpha_G = 0.057$ | 0.199 | $\alpha_G = 0.0097$ | 0.169 |
| $\mu_{\text{sil}} = 30.82$[d] | | $\mu_{\text{sil}} = 2.05$[d] | |

a – Likelihoods for 14 strains of *E. coli* (1407bp; see Table 1).

b – Pooled likelihoods for 6 *Drosophila simulans* and 12 *D. yakuba* strains (771bp; McDONALD and KREITMAN, 1991; pooling means that within-species log likelihoods are summed).

c – Base frequencies at 4-fold degenerate regular silent sites.

d – Locus-wide silent mutation rate scaled by $N_e$ (see text).

of twofold degenerate sites, can be combined to obtain maximum likelihood estimators for $\alpha_T, \alpha_C, \alpha_A$, and $\alpha_G$ (Table 2).

Given equation (1), the mean frequency of the base T at fourfold degenerate sites is $E(p_T) = \alpha_T/\alpha$ for $\alpha$ in equation (2). Thus the expected rate of transitions T $\rightarrow$ C at fourfold degenerate sites in a genetic locus is $N_4\alpha_T\alpha_C/(2\alpha)$, where $N_4$ is the number of fourfold degenerate regular codon positions in the locus. (The factor of two is because $\mu_C$ is the mutation rate to the base C per $N_e$ generations, while $\alpha_C = 2N_e\mu_C$ in equation (1).) Similarly, the mutation rate at pyrimidine twofold degenerate sites in the locus is $N_{2,TC}\,\alpha_T\alpha_C/(\alpha_T + \alpha_C)$, where $N_{2,TC}$ is the number of pyrimidine twofold degenerate regular codon positions. These considerations lead to a formula for the locus-wide silent mutation rate $\mu_{\text{sil}}$ (Table 2).

**Remarks.** The maximum likelihood method assumes that the distributions at different silent sites can be treated as independent. Recombination and gene conversion both help to insure the independence of site distributions. Independence can be tested by computing the significance of autocorrelations of the events monomorphic/polymorphic for adjacent silent sites. The first three autocorrelations are not significant for either the *E. coli* data in Table 1

nor the two ADH data sets in Table 2. Maximum likelihood theory uses independence only for the central limit theorem for the log likelihoods for the various terms, and so can tolerate some deviation from joint statistical independence.

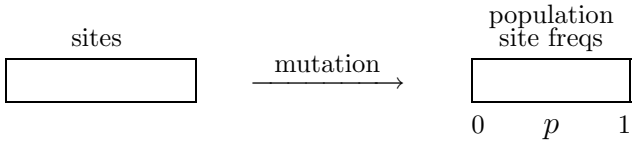
The methods that are used in this paper assume that each sample is a random sample from a panmictic population. If a sample contains some strains that are significantly different from the others, then model parameters will be estimated incorrectly. For example, the references quoted in Table 1 have 16 strains of *E. coli* at the *gnd* locus, but two strains (labeled r4 and r16) are as distant from the other *E. coli* strains as they are from *Salmonella*. The remaining *E. coli* strains have an estimated phylogeny with a more coalescent-like appearance. These two aberrant *E. coli* strains were excluded from the analysis. Similarly, the 13 *Salmonella* strains at the *PutP* locus in Table 1 were reduced to 8 strains.

## 4. Estimating Mutation *and* Selection Rates

We now discuss a model that will allow us to estimate the mutation rate $\mu$ and the relative selection rate $\gamma$ for mutants, both scaled by the haploid effective population size $N_e$ (SAWYER, 1994; HARTL, MORIYAMA, and SAWYER, 1994; *see also* SAWYER and HARTL, 1992).

This model is sensitive to saturation (repeated mutations at the same site), but should give reliable results if the estimate for $\mu_{\text{sil}}$ (the parameter $\mu$ for regular silent sites) is comparable to or greater than the more accurate estimate of $\mu_{\text{sil}}$ based on the Dirichlet density (1) of the previous section (which, however, assumes selective neutrality at silent sites). This model will be applied both for bases at regular silent sites and for amino acids at codon positions.

Consider a flux of mutations at the rate of $\mu$ per generation into the population. Each mutation changes one base at one site in one individual, and each new mutant base confers a relative selective advantage of $s = \gamma/N_e$ with respect to the current base. Most of the resulting new mutant alleles quickly go extinct by chance, but some survive to have appreciable base frequencies in the population:




Subsequent mutations are ignored at that site. Since we are assuming that all new mutant bases (or amino acids) are selectively equivalent, we can ignore mutation between mutant bases at the same site.

We now view the *population frequencies at polymorphic sites* for the surviving mutant bases as a *point process* of frequencies on $[0, 1]$. Under the usual diffusion approximation conditions (EWENS, 1979; ETHIER and KURTZ, 1986), this will be a Poisson point process with the expected density

$$\begin{aligned} 2\mu \frac{1 - e^{-2\gamma(1-p)}}{1 - e^{-2\gamma}} \frac{dp}{p(1-p)} \qquad & \text{for} \quad 0 < p < 1 \\ = 2\mu \frac{dp}{p} \qquad & \text{if } \gamma = 0 \text{ (i.e., no selection)} \end{aligned} \tag{3}$$

(See SAWYER and HARTL, 1992, for a sketch of the proof.) Note that the densities in (3) are *not* integrable at $p = 0$. This corresponds to the fact that the population contains a large number of rare mutants at any one time. The formula (3) was first derived by SEWALL WRIGHT (1938) as the transient distribution of the frequency of a single allele under selection and irreversible mutation.

Now suppose that we have a sample of $n$ aligned DNA sequences from this population. Let $N_k$ be the number of polymorphic sites that have $k$ bases different from the ancestral base at that site. Then the $N_k$ $(1 \le k \le n-1)$ are independent Poisson random variables with means

$$\begin{aligned} N(k, \mu, \gamma) &= 2\mu \int_0^1 \frac{1 - e^{-2\gamma(1-p)}}{1 - e^{-2\gamma}} \binom{n}{k} p^k (1-p)^{n-k} \frac{dp}{p(1-p)} \\ &= 2\mu \frac{1}{k} \qquad \text{if } \gamma = 0 \text{ (i.e., no selection)} \end{aligned} \tag{4}$$

(SAWYER and HARTL, 1992).

In practice, we will not be able to tell which base is the ancestral base by looking at the various bases at a polymorphic site. However, any polymorphic site has at most two bases in our model, of which *one* must be the ancestral base. Thus the number $X_k$ of polymorphic sites that have *either* $k$ or $n - k$ bases different from the ancestral base—or, equivalently, have $k$ of one type of base and $n - k$ bases of a second type—*is* observable. Here

$$\begin{aligned} X_k &= N_k + N_{n-k}, \qquad && 1 \le k < n/2 \\ &= N_{n/2}, && k = n/2 \end{aligned} \tag{5}$$

The variables $X_k$ are independent Poisson random variables with means

$$\begin{aligned} G(k, \mu, \gamma) &= N(k, \mu, \gamma) + N(n-k, \mu, \gamma), \qquad && 1 \le k < n/2 \\ &= N(n/2, \mu, \gamma), && k = n/2 \end{aligned} \tag{6}$$

for $N(k,\mu,\gamma)$ in (4). Since the $X_k$ are independent Poisson, the joint likelihood is

$$L(\mu,\gamma) = \prod_{k=1}^{(n+1)/2} e^{-G(k,\mu,\gamma)}\, \frac{G(k,\mu,\gamma)^{X_k}}{X_k!} \tag{7}$$

If a site actually has three or more different bases in a sample, choose $k$ so that $n-k$ is the number of representatives of one of the bases with the largest number of representatives in the sample. This is equivalent to assuming that either that base or the common ancestor of all the other bases is the ancestral base. Since there are likely to be relatively few sites with three or more different bases, this treatment is not likely to bias the analysis.

Since the expression $G(k,\mu,\gamma)$ in (6) can be written in the form $G(k,\mu,\gamma) = 2\mu J(k,\gamma)$ where $J(k,\gamma)$ does not depend on $\mu$, maximizing (7) is equivalent to the following procedure. First, set

$$\widehat{\mu}_{MLE} = \widehat{\mu}_{MLE}(\gamma) = \frac{X_{TOT}}{2J(\gamma)} \qquad \text{where} \quad X_{TOT} = \sum_{k=1}^{(n+1)/2} X_k$$

is the total number of polymorphic sites and $J(\gamma) = \sum_{k=1}^{(n+1)/2} J(k,\gamma)$. Then maximize

$$L(\gamma) = L(\widehat{\mu}_{MLE}(\gamma),\gamma) = C(X) \prod_{k=1}^{(n+1)/2} \left(\frac{J(k,\gamma)}{J(\gamma)}\right)^{X_k} \tag{8}$$

as a function of $\gamma$. In particular, finding $\widehat{\mu}_{MLE}$ and $\widehat{\gamma}_{MLE}$ can be reduced to a one-dimensional maximization.

Given a data set composed of $n$ aligned DNA sequences, we use the counts for polymorphic silent sites to estimate parameters $\mu_{\rm sil}$ and $\gamma_{\rm sil}$ for silent sites, and the counts for amino-acid polymorphic codon positions to estimate parameters $\mu_{\rm rep}$ and $\gamma_{\rm rep}$ for replacement amino acids. The conclusions for the two data sets of Table 1 are given in Table 3.

The scaled selection rate $\gamma_{\rm rep} = -3.66$ for *E. coli* in Table 3 corresponds to a selection rate of $s = -\gamma_{\rm rep}/N_e$ per generation against replacements, where $N_e$ is the effective population size of *E. coli*. We can estimate $N_e$ as follows. The value $\mu_{\rm sil} = 30.82$ in Table 2 corresponds to $N_{\rm sil} \times \mu N_e$, where $\mu$ is the mutation rate per site per generation and $N_{\rm sil}$ is the number of amino-acid monomorphic codon positions with twofold or fourfold degenerate regular silent sites. The 14 strains of *E. coli* in Table 2 have $N_{\rm sil} = 367$, and the estimate $\mu = 5 \times 10^{-10}$ per generation (OCHMAN and WILSON, 1987) implies $N_e = 1.7 \times 10^8$.

**Table 3: Joint estimates of the locus-wide mutation rate $\mu$ and the selection rate $\gamma$**

14 *E. coli* strains, *gnd* locus (1407p):

| | |
|---|---|
| $\mu_{\text{sil}} = 30.82$ | (neutral Wright model) |
| $\mu_{\text{sil}} = 33.57 \pm 5.50$[a] | |
| $\gamma_{\text{sil}} = -1.34 \pm 0.83^{**}$ [ac] | |
| $\mu_{\text{rep}} = 12.51 \pm 4.47$[b] | ($\gamma_{\text{sil}} \neq \gamma_{\text{rep}} : P = 0.029^{*}$) |
| $\gamma_{\text{rep}} = -3.66 \pm 2.24^{***}$ [bc] | |

8 *S. typhimurium* strains, *PutP* locus (1467bp):

| | |
|---|---|
| $\mu_{\text{sil}} = 41.05$ | (neutral Wright model) |
| $\mu_{\text{sil}} = 65.41 \pm 10.40$ | |
| $\gamma_{\text{sil}} = -2.43 \pm 0.96^{***}$ | |
| $\mu_{\text{rep}} = 7.63 \pm 3.34$ | ($\gamma_{\text{sil}} \neq \gamma_{\text{rep}} : P = 0.77$) |
| $\gamma_{\text{rep}} = -2.04 \pm 2.44$ | |

$* \; P < 0.05$ $\quad ** \; P < 0.01$ $\quad *** \; P < 0.001$

a – Estimated from base distributions at polymorphic regular silent sites.

b – Estimated from amino acid distributions at amino-acid polymorphic codon positions.

c – The ranges $\pm$ are 95% normal-theory confidence intervals, while P-values are for likelihood ratio tests against $\gamma = 0$.

This estimate of $N_e$ leads to $s = -\gamma_{\text{rep}}/N_e = 2.2 \times 10^{-8}$ per generation against replacements. Thus the average magnitude of selection per generation that acts against observed amino acid substitutions is quite small. One way in which such a small selection coefficient could be realized is if a substitution is selectively neutral in most environments, but disadvantageous in some rarely-encountered environments (Hartl, 1989).

The estimates $\mu_{\text{rep}} = 12.51$ (for 469 codons, or $2 * 469 = 938$ first and second codon position sites) but $\mu_{\text{sil}} = 30.82$ (for 367 codons) in Table 3 suggests that about one sixth of amino acid positions in *E. coli* are susceptible to a weakly-selected replacement. Sawyer, Dykhuizen, and Hartl (1987) estimated $s = 1.6 \times 10^{-7}$ against replacement amino acids in a similar model

assuming that *all* codon positions in *E. coli* in *gnd* were vulnerable to a weakly-selected amino-acid replacement. The earlier estimate is about seven times as large as the value $s = 2.2 \times 10^{-8}$ obtained above (with selection acting on about six times more amino acids), and the two estimates for $s$ are remarkably consistent.

The closeness of the estimates of $\mu_{\text{sil}}$ from the Poisson random field model to the estimates from Wright's Formula (2) suggests that saturation or repeated mutations at the same site do not have a significant effect on the estimates in Table 3. The fitted values for the numbers of polymorphic silent sites are quite close in both cases (Table 4). The fitted values for the counts for replacement amino acids resembled the observed counts in both cases, but had too many empty or near-empty cells to carry out a chi-square goodness-of-fit test.

**Table 4: Observed versus fitted values for the counts (5) for silent polymorphic sites**

14 *E. coli* strains, *gnd* locus (143 silent polymorphic sites): *

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Obs. $X_k$ | 60 | 31 | 16 | 10 | 14 | 7 | 5 |
| Est. $X_k$ | 60.97 | 28.08 | 17.59 | 12.76 | 10.26 | 9.02 | 4.32 |

8 *S. typhimurium* strains, *PutP* locus (152 silent polymorphic sites): **

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Obs. $X_k$ | 93 | 33 | 17 | 9 |
| Est. $X_k$ | 92.23 | 33.77 | 18.54 | 7.45 |

$*$ $P = 0.712$ (5 d.f.; $\mu_{\text{sil}} = 33.57,\ \gamma_{\text{sil}} = -1.34$)

$**$ $P = 0.796$ (2 d.f.; $\mu_{\text{sil}} = 65.41,\ \gamma_{\text{sil}} = -2.43$)

The goodness-of-fit test in Table 4 is a nested hypothesis test for $r$ Poisson variables $X_1, X_2, \ldots, X_r$. The means $\mu_k = E(X_k)$ are arbitrary in the larger model but satisfy $\mu_k = G(k, \mu_{\text{sil}}, \gamma_{\text{sil}})$ in the restricted model. If the restricted

model is true, then twice the logarithm of the ratio of the maximum likelihoods of the data under the two models has a $\chi^2$-distribution with $r-2$ degrees of freedom (Rao, 1973, p418). Note that we cannot use a standard $\chi^2$ cell test here since the sum of the counts is not constrained to have a preassigned value.

## References


Bisercic, M., J. Y. Feutrier, and P. R. Reeves (1991) Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. *J. Bacteriol.* 173, 3894–3900.

Dykhuizen, D. E., and L. Green (1991) Recombination in *Escherichia coli* and the definition of biological species. *J. Bacteriol.* 173, 7257–7268.

Ethier, S. N., and T. G. Kurtz (1986) *Markov Processes*. Wiley and Sons, New York.

Ewens, W. J. (1979) *Mathematical Population Genetics*. Springer-Verlag, New York.

Fu, Y.-X., and W.-Hs. Li (1993) Maximum likelihood estimation of population parameters. *Genetics* 134, 1261–1270.

Harris, H. (1966) Enzyme polymorphisms in man. *Proc. Royal Soc. London Ser. B* 164, 298–310.

Hartl, D. L. (1989) Evolving theories of enzyme evolution. *Genetics* 122, 1–6.

Hartl, D. L., and A. Clark (1989) *Principles of population genetics*, 2nd Ed. Sinauer Associates, Sunderland, MA.

Hartl, D. L., E. Moriyama, and S. A. Sawyer (1994) Selection intensity for codon bias. *Genetics* 138, 227–234.

Hartl, D. L., and S. A. Sawyer (1991) Inference of selection and recombination from nucleotide sequence data. *J. Evol. Biol.* 4, 519–532.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kingman, J. (1980) *Mathematics of Genetic Diversity*. CBMS-NSF Regional Conf. Ser. Appl. Math **34**, Soc. Ind. Appl. Math., Philadelphia.

Lewontin, R. C. (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press.

LEWONTIN, R. C., and J. L. HUBBY (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54, 595–609.

LI, W.-H., and D. GRAUR (1991) *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, MA.

LI, W.-H., C.-I. WU, and C.-C. LUO (1985) A new method of estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150–174.

MCDONALD, J. H., and M. KREITMAN (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351, 652–654.

NELSON, K., and R. K. SELANDER (1992) Evolutionary genetics of the proline permease gene (*PutP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. *J. Bacteriol.* 174, 6886–6895.

OCHMAN, H., and A. C. WILSON (1987) in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* eds Ingraham, J. L., Low, K. B., Magasanik, B., Neidhardt, F. C., Schaechter, M. and Umbarger, H. E. American Society of Microbiology Pubs.

RAO, C. R. (1973) *Linear statistical inference and its applications*, 2nd ed. John Wiley & Sons, New York.

SAWYER, S. A. (1994) Inferring selection and mutation from DNA sequences: The McDonald-Kreitman test revisited. *In* G. B. Golding (Ed.) *Non-Neutral Evolution: Theories and Data.* Chapman & Hall, New York, 77–87.

SAWYER, S. A., D. E. DYKHUIZEN, and D. L. HARTL (1987) Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Nat. Acad. Sci. USA* 84, 6225–6228.

SAWYER, S. A. and D. L. HARTL (1992) Population genetics of polymorphism and divergence. *Genetics* 132, 1161–1176.

TAJIMA, F. and M. NEI (1982) Biases of the estimates of DNA divergence obtain by the restriction enzyme technique. *J. Mol. Evol.* 18, 115–120.

WATTERSON, G. (1977) Heterosis or neutrality? *Genetics* 85, 789–814.

WILLS, C. (1973) In defense of naïve pan-selectionism. *Amer. Naturalist* 107, 23–34.

Wolfe, K., P. Sharp, and W.-H. Li (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337, 283–285.

Wright, S. (1938) The distribution of gene frequencies under irreversible mutation. *Proc. Nat. Acad. Sci. USA* 24, 253–259.

Wright, S. (1949) Adaption and selection, pp365–389 in *Genetics, Paleontology, and Evolution*, edited by G. Jepson, G. Simpson, and E. Mayr. Princeton Univ. Press, Princeton, N.J.