Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 6225–6228, September 1987
Genetics

Confidence interval for the number of selectively neutral amino acid polymorphisms

(*Escherichia coli/gnd* gene/molecular evolution/6-phosphogluconate dehydrogenase/sampling theory)

STANLEY A. SAWYER*†, DANIEL E. DYKHUIZEN*, AND DANIEL L. HARTL*‡

*Department of Genetics and †Department of Mathematics, Washington University, St. Louis, MO 63130

Communicated by Peter H. Raven, May 5, 1987 (received for review February 11, 1987)

ABSTRACT **A statistical approach to the analysis of DNA sequences has been developed, which provides a confidence interval estimate for the proportion of naturally occurring amino acid polymorphisms that are selectively neutral. When applied to the *gnd* gene coding for 6-phosphogluconate dehydrogenase in a sample of seven natural isolates of *Escherichia coli*, the method indicates that the proportion of observed amino acid polymorphisms that are selectively neutral is unlikely to be greater than 49% (upper 95% confidence limit). On the other hand, the observations are also consistent with a model in which all of the observed amino acid substitutions are mildly deleterious with an average selection coefficient approximating 1.6×10^{-7}. Various models for the distribution of configurations at silent sites are also investigated.**

Section 1. Introduction

The proportion of observed amino acid polymorphisms that are selectively neutral has been a central question in population genetics for over a decade (1, 2). It has so far defied resolution, in part because statistical tests of observed gene frequencies (3) and laboratory experiments (4) are too lacking in power to detect selection coefficients of the relevant magnitude. Kreitman (5) determined the nucleotide sequence of the alcohol dehydrogenase gene in 11 strains of *Drosophila melanogaster* and found one amino acid polymorphism and 13 silent nucleotide polymorphisms within the coding region. This result led him to conclude that most mutations that change amino acids in the alcohol dehydrogenase gene are harmful. Kimura (2, 6), using the frequencies of rare variant alleles to estimate the fraction of alleles P_{neut} that are selectively neutral among all newly arising mutations, estimated that P_{neut} ranges from 0.06 to 0.21 (mean = 0.14 ± 0.06).

An equally challenging problem is to estimate the fraction π_{neut} of observed amino acid polymorphisms that are selectively neutral. Although most newly arising mutations may be harmful, only a small proportion of harmful alleles survive long enough to become polymorphic in natural populations. A significant proportion of alleles that become polymorphic might therefore be expected to be selectively neutral or nearly neutral.

We have developed a method of estimating π_{neut}. The method has been applied to the sequences of nucleotides at positions 405–1172 in the *gnd* genes of seven natural isolates of *Escherichia coli* (numbered as in ref. 7). These nucleotides form codons 117–372 in the 468-amino acid polypeptide and contain 12 amino acid polymorphisms and 78 silent polymorphisms among the 256 codon positions. The amino acid polymorphisms are summarized in Table 1. All 12 amino acid polymorphisms have the configuration "6, 1" in the sample, which means that six strains code for the same amino acid at this position and one minority strain codes for a different amino acid. Although the identities of the minority amino acids are essentially random, the majority amino acids in these data form a nonrandom subset of all amino acids in the sequenced region of the enzyme ($P < 0.01$); that is, codons coding for certain amino acids are significantly more likely to be associated with amino acid polymorphisms.

Consensus codons can be classified according to the degeneracy of their nucleotide sites (9). Sites with 4-fold degeneracy code for the same amino acid when occupied by any nucleotide, and sites with 2-fold degeneracy code for the same amino acid when occupied by either of two nucleotides (usually either pyrimidine or either purine). The nucleotide sequences of the *gnd* genes contain 208 amino acid monomorphic codon positions in which the 3′ position is unambiguously either 2-fold or 4-fold degenerate§. Table 2 summarizes the sample configurations at the 3′ sites in these codons. The configuration "7, 0" represents sample monomorphism, "6, 1" represents six strains with the consensus nucleotide and one singleton, "5, 1, 1" represents five strains with the consensus nucleotide and two different singletons, and so on.

Among the silent polymorphisms in Table 2, which are presumably selectively neutral or nearly so, 34 out of 66 occur in "6, 1" configurations. In contrast, 12 out of 12 of the amino acid polymorphisms in Table 1 are the result of "6, 1" nucleotide configurations. If the amino acid polymorphisms were neutral, the probability that all of the nucleotide polymorphisms would be in "6, 1" configurations is approximately 0.0026 by means of the Fisher exact test for a 2×2 contingency table, which is statistically highly significant. If only 2-fold degenerate silent sites are used (with 24 polymorphisms), the result is still significant ($P \simeq 0.036$). In a different vein, the *a priori* probability that a neutral sample of size seven with two types in the infinite-alleles model has the configuration "6, 1" is 0.4762 independently of the mutation rate (10). The data in Table 1 are highly significant in this model as well, since $0.4762^{12} = 0.00014$. Both of these arguments implicitly assume that the codon positions are independent. Due to common ancestry of the strains, the codon positions may not be independent. However, contingency table tests can be applied without assuming linkage equilibrium. This is because, assuming selective neutrality, the conditional distribution of the codon configurations given the pedigree of the seven strains depends only on the mutational history of the strains since a common ancestor, so

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: P, level of statistical significance.

‡To whom reprint requests should be addressed.

§We excluded 31 leucine and arginine codons, which are ambiguous in the sense that a silent change in the first position can alter the degeneracy in the third position. Isoleucine codons were considered 2-fold degenerate since the ATA codon is rare in *E. coli* and did not occur in the seven strains.

Table 1. Distribution of amino acid polymorphisms

Codon position	Polymorphism*	Site change	Major codon†	Minority strain‡
117	Ala, Ser	1	004	RM70B
175	Ala, Ser	1	004	RM201
209	Asn, Ser	2	002	RM224
211	Thr, Ser	1	004	RM45E
216	Ala, Thr	1	004	RM217
294	Asp, Glu	3	002	RM70B
306	Pro, Arg	2	004	K12
308	Ala, Gly	2	004	RM224
313	Asp, Asn	1	002	RM217
315	Ala, Gly	2	004	RM70B
325	Leu, Gln	2	204	RM45E
350	Asp, Ala	2	002	RM45E

*All 12 amino acid polymorphisms have the configuration "6, 1" in the sample.
†Degeneracy of nucleotides in the consensus codon.
‡Strain containing the minority amino acid. Strains RM201 and RM217 were isolated from a domestic pig and a domestic goat, respectively, in Indonesia. Strains RM45E and RM202I were isolated from a Celebese black ape, RM70B from a lowland gorilla, and RM224 from a giraffe. All animals were in the Woodland Park Zoo, Seattle (21).

that the codon configurations are independent given the pedigree.

Although the silent sites in the seven strains do not show significant autocorrelation with respect to monomorphism versus polymorphism (see *Section 2*), there should be a small positive correlation between all pairs of silent sites unless ameliorated by intragenic recombination (see *Section 3*). On the other hand, the infinite-alleles model may not be appropriate for 2-fold or 4-fold degenerate silent sites, even if there were sufficient intragenic recombination to guarantee linkage equilibrium. We therefore develop a more appropriate model, from which relevant parameters are estimated and a confidence bound is calculated for the maximum proportion of the 12 amino acid polymorphisms that can be selectively neutral.

Section 2. Basic Models and Conclusions

Consider a population model of the Moran type (11), in which individuals die at a constant rate and are replaced by possibly mutated replicas of other individuals. We follow a particular nucleotide site with degeneracy k, where $k = 2$ or 4. The population frequencies in a large stationary Moranian population with k types and equal between-type mutation rates have a Dirichlet distribution (11–13). When $k = 4$, the Dirichlet distribution has density $[\Gamma(4\alpha)/\Gamma(\alpha)^4](x_1x_2x_3x_4)^{\alpha-1}$, where Γ is the gamma function and x_1, x_2, x_3, and x_4 are the frequencies of the four nucleotides. The parameter α includes the effects of mutation, recombination, and sampling drift. If $k = 2$, the Dirichlet distribution reduces to a β distribution. We call this model the Dirichlet model for definiteness.

Table 2. Sample configurations at 208 silent sites

Configuration	Degeneracy 2-fold	Degeneracy 4-fold
7, 0	92	50
6, 1	15	19
5, 2	5	7
4, 3	4	5
5, 1, 1		3
4, 2, 1		3
3, 3, 1		2
3, 2, 2		2
4, 1, 1, 1		1
3, 2, 1, 1		0
2, 2, 2, 1		0
Total sites	116	92
Polymorphic sites	24	42

Sample configurations in this model can be written $\gamma = (n_1 \ldots n_k)$, where $n_1 \geq \ldots \geq n_k \geq 0$ for $k = 2$ or 4 and n_i equals the number of strains carrying the ith most common nucleotide. Assuming r strains, the probability of γ given α is

$$P_r(\gamma|\alpha) = C(\gamma)\frac{\prod_{i=1}^{k}\alpha^{(n(i))}}{(k\alpha)^{(r)}}, \qquad [2.1]$$

for $C(\gamma) = r!k!/[n_1! \ldots n_k!d(0)! \ldots d(r)!]$, where $d(j)$ is the number of indices i such that $n_i = j$ and $x^{(k)} = x(x + 1) \ldots (x + k - 1)$ (14). A maximum likelihood estimate for α was calculated from the observed nucleotide configurations at the 208 sites that are 2-fold and 4-fold degenerate. Based on Table 2, $\alpha = 0.104$ with 95% confidence interval (0.078, 0.130). The predicted sample distributions fit the data ($\chi^2 = 8.4$ with five degrees of freedom, and $P \simeq 0.13$). Maximum likelihood procedures generally assume that the data are independent, but they are relatively insensitive to mild autocorrelation in the data. The Pearson autocorrelation for monomorphism for pairs of silent sites from the 208 positions are 0.136, −0.124, −0.047, and −0.028, respectively, for pairs that are physically 1–4 codon positions apart. None of these autocorrelations is statistically significant. On the other hand, the model predicts a Pearson autocorrelation of 0.039 between all pairs of 2-fold degenerate silent sites and 0.082 between all pairs of 4-fold degenerate sites (see *Section 3*), based on seven strains with $\alpha = 0.104$. These correlations might be considered medium to large if they occurred between all pairs of silent sites. However, even a few intragenic recombination events since the common ancestor of the seven strains would weaken the overall average pairwise correlation considerably.

Suppose that a fraction π_{neut} of all amino acid polymorphisms are selectively neutral, with the rest perhaps showing the effects of selection. Let $c(\alpha, k)$ be the probability that a selectively neutral k-fold degenerate polymorphic site has a "6, 1" sample configuration. Recall that α is a time constant multiplied by the base mutation rate and so should be the same at 2-fold and 4-fold degenerate sites. The 95% confidence interval for α estimated from Table 2 yields 95% confidence intervals of (0.449, 0.459) for $c(\alpha, k)$ for $k = 2$ and (0.359, 0.400) for $k = 4$. Thus, with 97.5% confidence, $c(\alpha, k) \leq c = 0.459$ in all cases. Given π_{neut}, a randomly chosen amino acid polymorphism will be "6, 1" with probability bounded by $(1 - \pi_{neut} + c\pi_{neut})$, with 97.5% confidence. Similarly, m independent amino acid polymorphisms will all be of type "6, 1" with probability bounded by $(1 - \pi_{neut} + c\pi_{neut})^m$, with 97.5% confidence. Solving for π_{neut} in $(1 - \pi_{neut} + c\pi_{neut})^m = 0.025$ leads to $[0, (1 - 0.025^{1/m})/(1 - c)]$ as a 95% confidence interval for π_{neut}, or (0, 0.489) for $c = 0.459$ and $m = 12$. That is, the number of selectively neutral amino acid polymorphisms in the sample of 12 may be as few as zero but, with 95% confidence, can be asserted to be no greater than six.

On the other hand, the data are also consistent with a model in which no amino acid polymorphisms are selectively neutral. Assume that one particular nucleotide at each nondegenerate first or second position of a codon is favored, while the other three nucleotides are equally deleterious. Table 1 shows 10 polymorphisms at such sites out of 450 in the seven strains. When selection is incorporated into the

Moran model, the population frequency x of the favored nucleotide has the density $Ce^{\sigma x}x^{\alpha-1}(1-x)^{3\alpha-1}$, where C is a normalization constant and σ is a measure of selection (11). Using the value of α estimated from the silent polymorphisms, a maximum likelihood estimate of σ is $\sigma = 99.3$ with 95% confidence interval (36.6, 161.9). The ratio $\sigma/\alpha = 954$. The fitted model is consistent with no amino acid polymorphisms having a configuration more complex than "6, 1" ($P \simeq 0.69$). With these values of the parameters, the probability that seven strains would be monomorphic for a disfavored nucleotide is 7×10^{-15}.

The parameter α in the Dirichlet distribution can be written as $\mu/3\lambda$, where μ is the rate of nucleotide mutation and λ is the rate, per pair of organisms, at which the first organism is replaced by a replica of the second by population regulation or else its *gnd* gene is replaced by recombination. Selection is modeled by letting $\lambda(1 + \sigma/R)$ be the rate at which favored organisms replace disfavored ones relative to the rate λ at which disfavored organisms replace favored ones, where R is the neighborhood size. If disfavored organisms occur with frequency q, the estimated rate of increase in the favored genotypes is $qR(\lambda\sigma/R) = q(\mu\sigma/3\alpha) = 1.6 \times 10^{-7}q$, assuming $\sigma/\alpha = 954$ and $\mu = 5 \times 10^{-10}$ per generation (15). The constant 1.6×10^{-7} corresponds to the conventional selection coefficient. Thus, the actual magnitude of selection that needs to be invoked against amino acid substitutions in order to account for the exclusively "6, 1" sample configurations is quite small.

The inference that at least some *gnd*-encoded amino acid polymorphisms are subject to natural selection is supported by chemostat studies on gluconate medium (4, 16)¶. The estimated magnitude of selection under natural conditions is very small, but nevertheless significant in light of the assumed large effective population number in *E. coli* (17, 18). In organisms with a smaller effective population number and with alleles having selection coefficients comparable in magnitude to those of the *gnd* gene in *E. coli*, the alleles might have an appreciable probability of becoming fixed through random genetic drift. Therefore, our results with the *gnd* gene are consistent with Ohta's model of molecular evolution, which is driven by the random drift of slightly deleterious alleles (8, 19).

Section 3. Additional Models and Conclusions

The Bernoulli Model. An alternative to the equilibrium (neutral) Moran model is a "star" or "palmetto" phylogeny, in which the seven strains have had no recombination or sampling drift events since diverging from a common ancestor. In this model, let p be the probability that a given site in a given strain has undergone one or more mutations since the common ancestor. For mathematical convenience, we assume that when a nucleotide mutates, it mutates to any of k bases (including itself) with equal probability. Thus a mutation randomizes the nucleotide at that site. Assuming neutrality, the nucleotide will be identical with the ancestral nucleotide at that site with probability $1 - p + (p/k)$, and to any other nucleotide with probability p/k. The configuration probabilities for r strains are

$$P_r(\gamma|p) = C(\gamma)\sum_{j=1}^{r}\frac{d(j)}{k}[1 - p + (p/k)]^j(p/k)^{r-j}, \quad [3.1]$$

for $C(\gamma)$ and $d(j)$ in Eq. **2.1**. However, the silent site data in Table 2 are highly inconsistent with the configuration probabilities in Eq. **3.1** ($P < 10^{-11}$, with maximum likelihood estimator for p). Thus, the seven strains show strong evidence for multiple divergences or recombination since their common ancestor.

The Purine–Pyrimidine Model. The Dirichlet model assumes that any nucleotide is equally likely to mutate to any other nucleotide. However, there is evidence that purines are more likely to mutate to purines, and pyrimidines are more likely to mutate to pyrimidines (9). We consider a model in which transition mutations (i.e., from a purine to the other purine, or a pyrimidine to the other pyrimidine) occur at the rate $u + v$, while transversion mutations (i.e., from a purine to a particular pyrimidine) occur at the rate u. Thus the total transversion rate is $2u$, and the total substitution rate is $3u + v$. Purines and pyrimidines must now be distinguished in sample configurations. Since the 2-fold silent sites in Table 2 involve either two purines or two pyrimidines, their configurations are the same as in *Section 2*. Fourfold configurations can be written $\gamma = (n_1n_2n_3n_4)$, where $n_1 \geq n_2$, $n_3 \geq n_4$, and either $n_1 > n_3$ or $n_1 = n_3$, $n_2 \geq n_4$. Examples would be (6100) (i.e., six strains monomorphic for one of the purines and one strain with the other purine) and (6010) (i.e., six strains with one purine and one strain with a pyrimidine). As in *Section 1*, a nonparametric argument shows that transition mutations must be more common than transversion mutations but does not provide a good quantitative estimate. Specifically, if purines and pyrimidines are distinguished and if transition and transversion mutations are equally likely, then one-third of all 4-fold silent "6, 1" configurations will be (6100) and two-thirds will be (6010). This is because, if the consensus base is a purine, there are twice as many pyrimidines as the remaining purine. However, the 19 "6, 1" configurations in Table 2 consist of 11 occurrences of (6100) and 8 of (6010). This is a significant discrepancy from a one-third to two-thirds proportion ($P \simeq 0.023$). Because one can make the argument conditional on the specified pedigree of the seven strains, this argument does not require linkage equilibrium or intragenic recombination since the common ancestor. Thus, the data show a significant difference between purine–purine and purine–pyrimidine mutation rates.

We now estimate the ratio $(u + v)/2u$ of the transition rate to the total transversion rate. The probabilities of configurations at silent sites will depend on two parameters, $\alpha = u/\lambda$ and $\beta = v/\lambda$, for λ in *Section 2*. In general, let $P_r(n_Tn_Cn_An_G)$ be the probability that r strains have n_T Ts, . . . at a given 4-fold degenerate site. Then, using the same type of arguments as in reference 13 (pages 24–27),

$$P_r(n_Tn_Cn_An_G) = \frac{\beta}{r(4\alpha + \beta + r - 1)} \times [(n_C + 1)P_r(n_T - 1, n_C + 1, n_A, n_G) + 3 \text{ other terms}] + \frac{1}{r(4\alpha + \beta + r - 1)}[(\alpha + n_T - 1)P_{r-1}(n_T - 1, n_C, n_A, n_G) + 3 \text{ other terms}]. \quad [3.2]$$

In principle, P_r can be found by induction by solving Eq. **3.2** beginning with $P_1(1000) = 1$. However, although there are only 20 configurations for $r = 7$, solving Eq. **3.2** as written would involve inverting a 120×120 matrix. We use a method (both here and once more below) that finds the configuration probabilities $P_r(\gamma)$ directly from Eq. **3.2** and does not require that the combinatorial coefficients that relate $P_r(n_T\ldots)$ to $P_r(\gamma)$ be known. Assuming $P_{r-1}(\gamma_0)$ is known for $(r-1)$-strain configurations γ_0, we first let $D_r(\gamma)$ be the sum of all terms in the second sum in Eq. **3.2** with $P_{r-1}(n_T\ldots)$ replaced by $P_{r-1}(\gamma_0)$ where $n_T\ldots$ has the configuration γ_0 and γ corresponds to γ_0 with one additional type i nucleotide ($1 \leq i \leq 4$).

¶It has not been determined whether the selective effects in chemostats result from structural or regulatory differences among the *gnd* genes.

Second, for r strain configurations γ_1 and γ_2, let $A_{rr}(\gamma_1, \gamma_2)$ be the sum of all integers n_i for $\gamma_2 = (n_1, \ldots, n_4)$ and $1 \leq i \leq 4$ such that a transition mutation of one type i nucleotide converts γ_2 to γ_1. Then, in matrix notation, Eq. **3.2** implies $P_r = [I - (\beta/r(4\alpha + \beta + r - 1))A_{rr}]^{-1}D_r$, where I is the identity matrix. Maximum likelihood estimators can then be found for α and β for data analogous to Table 2, using this procedure to find the vector P_7 at each iteration step. The largest matrix that must be inverted is 20×20.

Even with purines and pyrimidines distinguished, the purine–pyrimidine model fits the configurations at the 208 silent sites better than does the Dirichlet model of *Section 2* ($\chi^2 = 5.3$ with four degrees of freedom; $P \simeq 0.26$). As foreshadowed by the nonparametric argument, the purine–pyrimidine model is rejected if $\nu = \beta = 0$ ($\chi 2 = 11.2$ with five degrees of freedom; $P \simeq 0.049$). The maximum likelihood estimates are $\alpha = 0.0618$ and $\beta = 0.0591$. The estimated ratio of transition to total transversion rates is $(u + v)/2u = (\alpha + \beta)/2\alpha = 0.978$ with 95% confidence interval (0.431, 2.672). This is consistent with a published estimate of 1.49 (9).

Correlations Between Silent Sites. The Pearson correlation for monomorphism vs. polymorphism at silent sites is $\rho = (p_2 - p^2)/[p(1 - p)]^{1/2}$ where p is the probability of monomorphism at one site and p_2 is the probability that two tightly linked sites are both monomorphic. In general, neutral configuration sampling theory at two sites or loci is relatively undeveloped (see however ref. 20), even with tight linkage. If back mutation is neglected, the Ewens sampling formula (10) with $\theta = \alpha(k - 1)$ at one site and $\theta = \alpha(2k - 2)$ at two sites gives $\rho = 0.0497$ ($k = 2$) and $\rho = 0.0997$ ($k = 4$), for $\alpha = 0.104$. There are k^2 nucleotide pairs at two sites; each nucleotide pair mutates to (and back-mutates from) $2k - 2$ other nucleotide pairs at the same rate α. A Dirichlet approximation that takes some back mutation into account calculates the probability of joint monomorphism by using Eq. **2.1** with k replaced by $2k - 1$. This approximation gives $\rho = 0.0459$ ($k = 2$) and $\rho = 0.0946$ ($k = 4$). The exact solution is more difficult and takes into account all possible configurations of nucleotide states at two sites. There turns out to be a vast number ($\simeq 560$) of distinguishable configurations for two sites in $r = 7$ strains with $k = 4$ nucleotides at each site but only 30 for $k = 2$. The probability of monomorphism for two tightly linked sites was calculated using the method of Eq. **3.2** with iteration in Eq. **3.2** rather than matrix inversion for $k = 4$ and $r \geq 6$. The exact Pearson correlation between pairs of tightly linked sites is $\rho = 0.0393$ ($k = 2$) and $\rho = 0.0822$ ($k = 4$) when $\alpha = 0.104$, so in this case the approximations are reasonably good.

We are grateful to Louis Green for expert DNA sequencing and Roger Milkman and Richard E. Wolf, Jr., for providing bacterial strains. We would also like to thank Joseph Felsenstein for pointing out the appropriateness of a purine–pyrimidine model as well as Andrew Clark and Motoo Kimura for helpful comments on the manuscript. This work was supported in part by National Science Foundation Grant DMS-8504315 (S.A.S.) and National Institutes of Health Grant GM30201 (D.L.H. and D.E.D.).

1. Lewontin, R. C. (1974) *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, New York).
2. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.)
3. Watterson, G. (1978) *Genetics* **88,** 171–179.
4. Hartl, D. L. & Dykhuizen, D. E. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 6344–6348.
5. Kreitman, M. (1983) *Nature (London)* **304,** 412–417.
6. Kimura, M. (1983) *Mol. Biol. Evol.* **1,** 84–93.
7. Nasoff, M. S., Baker, H. V., II & Wolf, R. E., Jr. (1984) *Gene* **27,** 253–264.
8. Ohta, T. (1974) *Nature (London)* **252,** 351–354.
9. Li, W.-H., Wu, C.-I. & Luo, C.-C. (1985) *Mol. Biol. Evol.* **2,** 150–174.
10. Ewens, W. (1972) *Theor. Pop. Biol.* **3,** 87–112.
11. Moran, P. (1962) *The Statistical Processes of Evolutionary Theory* (Clarendon, Oxford).
12. Wright, S. (1949) in *Genetics, Paleontology, and Evolution*, eds. Jepson, G., Simpson, G. & Mayr, E. (Princeton Univ. Press, Princeton, NJ), pp. 365–389.
13. Kingman, J. F. C. (1980) *Mathematics of Genetic Diversity*, Regional Conf. Ser. Appl. Math. (Soc. Ind. Appl. Math., Philadelphia), Vol. 34.
14. Watterson, G. (1977) *Genetics* **85,** 789–814.
15. Ochman, H. & Wilson, A. C. (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. & Umbarger, H. E. (Am. Soc. Microbiol., Washington, DC).
16. Dykhuizen, D. E. & Hartl, D. L. (1980) *Genetics* **96,** 801–817.
17. Milkman, R. (1975) in *Isozymes*, ed. Markert, C. L. (Academic, New York), Vol. 4, pp. 271–285.
18. Maruyama, T. & Kimura, M. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 6710–6714.
19. Ohta, T. (1973) *Nature (London)* **246,** 96–98.
20. Hedrick, P. & Thomson, G. (1986) *Genetics* **112,** 135–156.
21. Milkman, R. & Crawford, I. P. (1983) *Science* **221,** 378–379.