J Mol Evol (1995) 41:449–456

JOURNAL OF MOLECULAR EVOLUTION

Synonymous Substitution-Rate Constants in *Escherichia coli* and *Salmonella typhimurium* and Their Relationship to Gene Expression and Selection Pressure

Otto G. Berg, Mattias Martelius

Department of Molecular Biology, University of Uppsala Biomedical Center, Box 590, S-75124 Uppsala, Sweden

Received: 24 November 1994 / Accepted: 28 February 1995

Abstract. Based on the differences in synonymous codon use between *E. coli* and *S. typhimurium,* the synonymous substitution rates can be estimated. In contrast to previous studies on the substitution rates in these two organisms, we use a kinetic model that explicitly takes the selection bias into account. The selection pressure on synonymous codons for a particular amino acid can be calculated from the observed codon bias. This offers a unique opportunity to study systematically the relationship between substitution-rate constants and selection pressure. The results indicate that the codon bias in these organisms is determined by a mutation-selection balance rather than by stabilizing selection. A best fit to the data implies that the mutation rate constant increases about threefold in genes at low expression levels relative to those that are highly expressed.

Key words: Synonymous codon usage — Synonymous substitution rates — Genome evolution — Mutation rates

Introduction

The genomic sequence divergence between two organisms can be used to estimate their evolutionary distance. With suitable models, the sequence divergence can also be used to calculate the nucleotide substitution rates. Most commonly, synonymous nucleotide changes—i.e., changes in coding sequences that do not lead to any change in amino acid—have been considered and treated as neutral. However, in many organisms, the choice between synonymous codons is clearly not neutral. This is evidenced by the strong preference for a particular codon choice among synonymous possibilities, the codon bias. In some organisms, *E. coli* and *S. typhimurium* among them, the codon bias becomes stronger in genes with a higher expression level (Ikemura 1981, 1985; Sharp and Li 1987). This correlation suggests that here is a selective advantage determined in some way by translational efficiency to use particular codons in genes that are translated often. This is supported also by the fact that in these organisms the preferred codons are read by major tRNAs (Ikemura 1981, 1985).

Homologous genes in *E. coli* and *S. typhimurium* have very similar codon bias (Sharp and Li 1987; Sharp 1991), which suggests that the selection pressure on synonymous codon choices has remained roughly the same in the two organisms since their separation; this is also supported by the present analysis. In each gene, the codon bias is the same throughout, except for the first 50–100 codons or so (Bulmer 1988a; Eyre-Walker and Bulmer 1993). Therefore, it seems reasonable that the selection pressure on synonymous codons is the same over the whole gene, except at its beginning. We will make the further assumption that genes with the same codon bias are under the same selection pressure on synonymous codons.

Correspondence to: O.G. Berg

Previous studies on the synonymous substitution rate in *E. coli* and *S. typhimurium* (Sharp and Li 1987; Sharp 1991) found a very strong negative correlation between this rate and the codon bias, indicating that the substitution rate decreases in genes at high selection pressure (high codon bias). However, the substitution rate used in these studies, the parameter K_s as defined by Li et al. (1985), is an estimate of the average number of synonymous changes per site. This number says very little about the actual substitution rate constants from one nucleotide to another for the following reasons. First, it is based on Kimura's (1980) two-parameter model for the kinetics of nucleotide exchange. This model is not applicable for sequences that have a codon or nucleotide bias, and in fact it predicts simply that the synonymous codons tend toward equal frequencies. Second, in the calculation of K_s one takes the average overall synonymous changes, regardless of what codons are involved. In this way, one gets an estimate for a single synonymous substitution rate with a small statistical variance. However, the price for this small variance is that much of the interesting information has simply been averaged out.

To distinguish this average rate of change, often referred to as the synonymous substitution rate, from the rates of substitution of a certain nucleotide for another, in the following we will refer to the latter as *substitution rate constants* in analogy with the rate constants of chemical kinetics. In the present communication we will study these synonymous substitution-rate constants for twofold degenerate sites with a model that includes the selection pressure, i.e., assuming that the rate of substitution to a preferred codon is different from the rate when a preferred codon is replaced. Furthermore, we will also take into account that the synonymous substitution rate constants for different amino acids may be different. Since the codon bias for each amino acid in each gene provides a measure of the selection pressure, this offers a rare opportunity to study systematically the relationship between substitution rate constants and selection pressure.

Data

Gene sequences were taken from the EMBL data bank. The following sequences were used from genes at very high expression level (VH): *hup*A, *omp*C, *omp*A, *rpl*L, *tuf*A, *gap*A, *csp*A, *fus*A; genes at high expression (H): *gln*A, *rpo*B, *rpl*J, *crr, rps*G, *env*m, *gly*A, *adk, omp*H; genes at medium-high expression (MH): *pts*G, *pts*I, *crp, hup*B, *prs*A, *pur*H, *pur*D, *kat*G, *mhd, gnd;* genes at medium-low expression (ML): *cys*K, *gdh*A, *ara*A, *his*F, *his*G, *his*H, *car*A, *uvr*A, *met*J, *par*C, *nrd*A; genes at low expression (L): *aro*A, *trp*A, *trp*B, *trp*E, *ara*D, *his*A, *his*B, *his*C, *his*IE, *mut*S, *pho*E, *ddl*A, *liv*F, *pab*B, *cys*H, *lex*A, *met*F; genes at very-low expression (VL): *env*Z, *pab*A, *che*A, *che*R, *che*Z, *che*Y, *btu*B, *ara*C, *icl*R, *ada, liv*G, *met*R, *cys*M, *dna*Q. The genes were placed in the respective groups as defined by Bulmer (1988b) according to their level of codon bias (codon adaptation index).

Model

Codon Bias. To clarify the assumptions and notations, let us briefly consider the kinetics of synonymous change for a twofold degenerate amino acid. Similar two-state models for DNA evolution have been considered, for instance, by Sueoka (1962). Let M and m denote the major and minor codon, respectively. These differ by a transition in the third position. Assume that the rate constant for replacing m by M in the population is α_1 and for replacing M by m is α_2. Then the rate of change for the probability, P_M, that the amino acid at a particular site is coded for by the major codon is

$$\frac{dP_M}{dt} = \alpha_1(1 - P_M) - \alpha_2 P_M \tag{1}$$

After a long time the equilibrium probabilities are

$$P_M^{eq} = \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{B}{1+B} \tag{2}$$

$$P_m^{eq} = 1 - P_M^{eq} = \frac{1}{1+B} \tag{3}$$

where

$$B = \frac{\alpha_1}{\alpha_2} = \frac{P_M^{eq}}{P_m^{eq}} \tag{4}$$

is a measure of the codon bias for the particular amino acid under consideration. Thus the codon bias B is here defined as the ratio of the probabilities for major and minor codons. The mutation-selection balance at equilibrium is given by eqs. (2) and (3) as $\alpha_1 P_m^{eq} = \alpha_2 P_M^{eq}$, so that the smaller value of α_2 is exactly compensated by the larger value for the probability of a major codon, P_M^{eq}. The equations assume that synonymous codons at different positions are replaced independently and with the same substitution rate constants. If we consider groups of sites with the same selection for codon bias, the probabilities P_m^{eq} and P_M^{eq} will correspond to fractions of observed codon usage.

The average rate of synonymous substitution at a twofold degenerate site can be expressed as

$$\alpha_1 P_m^{eq} + \alpha_2 P_M^{eq} = 2\left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2}\right)^{-1} \tag{5}$$

which is the harmonic mean of α_1 and α_2. Thus, the average rate of change at a certain site is dominated by the slow step, α_2.

Accumulation of Differences. When comparing homologous sequences from two different organisms, one can only observe the differences and similarities between them. If P_{Mm} denotes the probability that one sequence has a major codon and the other a minor one at a certain site, and if P_{MM} and P_{mm} indicate the probabilities that both use major and minor, respectively, their rate of change would be determined by

$$\frac{dP_{Mm}}{dt} = 2\alpha_1 P_{mm} + 2\alpha_2 P_{MM} - (\alpha_1 + \alpha_2) P_{Mm} \quad (6)$$

$$\frac{dP_{MM}}{dt} = \alpha_1 P_{Mm} - 2\alpha_2 P_{MM} \quad (7)$$

$$\frac{dP_{mm}}{dt} = \alpha_2 P_{Mm} - 2\alpha_1 P_{mm} \quad (8)$$

These equations assume that the substitution rate constants are the same in the two organisms. The probabilities are normalized, $P_{Mm} + P_{MM} + P_{mm} = 1$. The initial condition is $P_{Mm}(0) = 0$ since the sequences were the same at the time of separation of the two organisms. Assuming $B = \alpha_1/\alpha_2$, as before, the solution to the differential equations can be written

$$P_{Mm}(t) = \frac{2B}{(1+B)^2}(1 - \exp(-2(\alpha_1 + \alpha_2)t)) \quad (9)$$

Eq. (9) expresses the gradual randomization as to which codon is where in the two sequences. After a long time the differences are saturated (equilibrated) such that the probability that one sequence has a major codon while the other has a minor one at a certain site can be determined as the product of two independent events: $P_{Mm}(t = \infty) = 2P_M^{eq}P_m^{eq} = 2B/(B+1)^2$. The time scale in the relaxation toward saturation is determined by the relaxation rate $2(\alpha_1 + \alpha_2)$. In contrast to the average rate of change, eq. (5), this is dominated by the fast step, α_1. Only when $\alpha_1 = \alpha_2$, as tacitly assumed in the calculation of the K_s parameter, will the average rate of change correspond to the relaxation rate.

It should be noted that eq. (9) holds for the remaining synonymous positions also if there have been a number of nonsynonymous changes, as long as these changes occur with the same probability for both of the synonymous codons. The model does not account for synonymous positions that have appeared via an intermediate with a nonsynonymous change; these are expected to be very few in this data set. The relation (9) is valid also if the substitution rates are not the same in the two organisms as long as the codon bias is the same. In this case α_1 and α_2 in the equations would correspond to the arithmetic mean of their values in the two organisms.

In the comparison between homologous sequences from *E. coli* and *S. typhimurium*, we consider groups of genes with similar codon bias. Then we look at positions where there is no amino acid difference and count the fraction of synonymous codon differences (P_{Mm}) for each amino acid considered and in each group. For the expected equilibrium value of the codon bias for that amino acid and in that group, B from eq. (4), we use the value from the codon distribution from both organisms. However, since these codon distributions are for the entire lengths of the genes (including the beginning where codon bias is smaller), this introduces a small systematic underestimate of B. When B is defined, the substitution rate from a minor to a major codon can be calculated from eq. (9) as

$$2\alpha_1 t = \frac{-B}{B+1}\ln\left(1 - \frac{(B+1)^2}{2B}P_{Mm}(t)\right) \quad (10)$$

where t is the time since separation. The reverse rate, α_2, follows from eq. (4) as $\alpha_2 = \alpha_1/B$.

For comparison, in Kimura's (1980) two-parameter model for a two-codon amino acid, where $\alpha_2 = \alpha_1 = \alpha$, one would use $B = 1$ in eq. (10) to calculate the average synonymous substitution rate (the average number of changes):

$$K_S^{2P} = 2\alpha t = -\frac{1}{2}\ln(1 - 2P_{Mm}(t)) \quad (11)$$

In contrast, the average number of changes for sites with bias different from $B = 1$ is given by eq. (5) as:

$$K_S = -\frac{2B}{(B+1)^2}\ln\left[1 - \frac{(B+1)^2}{2B}P_{Mm}(t)\right] \quad (12)$$

Only at very short times will use of eq. (11) give a reasonable estimate for the average number of changes as given in eq. (12) for arbitrary bias. After long times, when the changes approach saturation so that $P_{Mm} = 2B/(B+1)^2$ from eq. (9), use of the two-parameter model would give an estimate for the separation time and the substitution rate which is totally unrelated to time and rates and determined only by the codon bias. The relationship between the real average number of changes and that estimated from the two-state model for various degrees of codon bias and various degrees of degeneracy is discussed in a separate communication (Berg 1995a).

Mutation and Fixation. In a model where mutations occur randomly and are fixed in the populations via genetic drift, one expects the substitution rate constants to be determined from the mutation rate and fixation probability as

$$\alpha_1 = \frac{u_1 N_e s}{1 - \exp(-N_e s)} \quad (13)$$

$$\alpha_2 = \frac{u_2 N_e s \exp(-N_e s)}{1 - \exp(-N_e s)} \quad (14)$$

$$B = \frac{P_M^{eq}}{P_m^{eq}} = \frac{\alpha_1}{\alpha_2} = \frac{u_1}{u_2}\exp(N_e s) \quad (15)$$

These relations are for a haploid organism and differ from the more commonly used versions for a diploid (e.g., Li and Graur 1991) by some factors of 2. u_1 and u_2 are the mutation rate constants, i.e., the rates with which the corresponding mutations appear in an individual cell. The substitution rate constant is the product of the mutation rate and the probability of fixation in the population. N_e is the effective population size, and s is the selective advantage of the major codon relative to the minor.

Bacterial populations are usually very large and the time it would take for a counterselected (even weakly) mutation to become fixed is exceedingly long. Thus it is more likely that the synonymous codon changes, which are under weak selection or counterselection, are fixed via hitchhiking with a strongly selected variant. It can be shown (Berg 1995b) that the effective substitution rate constant for a weakly selected variant via hitchhiking would be approximately (except for $N_e s \gg 1$)

$$\alpha_1 = u_1 \frac{\exp(N_e s) - 1}{N_e s} \quad (16)$$

The ratio of the substitution rate constants, eq. (15), remains the same as for the case without hitchhiking. In this relation the effective population size is determined by the average time between successive selective sweeps where a strongly selected variant takes over the population.

Table 1. B values

	VH	H	MH	ML	L	VL
lys[a]	6.7	3.4	2.8	2.6	3.8	3.3
glu[a]	3.9	3.1	2.6	2.1	2.0	1.7
gln[b]	16	5.9	3.3	2.2	2.4	2.4
tyr[c]	3.1	2.5	1.6	1.1	0.83	0.54
his[c]	3.1	4.5	1.5	1.2	1.1	0.71
asn[c]	25	7.3	4.8	2.0	1.5	1.7
asp[c]	2.5	1.5	0.84	0.71	0.57	0.60
phe[c]	5.9	2.2	1.6	1.2	0.76	0.53

[a] Codon bias for A over G
[b] Codon bias for G over A
[c] Codon bias for C over T

Table 2. P_{Mm} values

	VH	H	MH	ML	L	VL
lys	0.033	0.093	0.17	0.23	0.24	0.34
glu	0.081	0.088	0.19	0.23	0.35	0.32
gln	0.013	0.058	0.18	0.20	0.26	0.23
tyr	0.22	0.21	0.26	0.32	0.53	0.35
his	0.054	0.19	0.32	0.25	0.42	0.36
asn	0.054	0.063	0.16	0.25	0.33	0.37
asp	0.20	0.21	0.28	0.27	0.30	0.32
phe	0.072	0.23	0.19	0.27	0.33	0.35

Results

We have looked at eight of the nine twofold degenerate amino acids for a number of genes in each of the six codon-bias (or expression-level) groups defined by Bulmer (1988b). The study excludes cysteine because of its small numbers of occurrence. The bias, B, was calculated for each amino acid in each group (Table 1) by taking the ratio of the number of major codons and the number of minor codons in both organisms. The fraction of sites for each amino acid, P_{Mm}, that have different synonymous codon choices in the two organisms is listed in Table 2.

Using eq. (15) and the observed codon bias, B (Table 1), the selection force ($N_e s$) for each amino acid in each group can be calculated as

$$N_e s = \ln(B) - \ln(u_1/u_2) \tag{17}$$

The twofold degenerate amino acids studied here involve five cases of changes between T and C and three cases of A–G exchange. As gene expression becomes lower, the codon bias should approach the mutational bias, u_1/u_2. We assume that the mutational bias corresponds to the bias observed in the genes of the lowest expression level (VL); thus $u_1/u_2 = B_{VL}$. This can be justified by the fact that Bulmer (1990) finds the same patterns of bias in the transcribed and nontranscribed strands of these genes. Furthermore, the results are not very sensitive to this choice. Using eq. (10) the substitution rate constants can be calculated from P_{Mm} and B. In Fig. 1 is plotted $2\alpha_1 t$ vs. the selective force $N_e s = \ln(B/B_{VL})$ for the two groups with the highest expression. The results can be fitted fairly well to the theoretical relations, eq. (13) or (16). In the lower groups (data not shown), the range of $N_e s$ values is small and, as also expected from eq. (13) or (16), the substitution rates are fairly constant, corresponding to the mutation rate $2u_1 t$.

Another way of looking at the results in Fig. 1 is to calculate the mutation rate constant ($2u_1 t$) for each amino acid in each gene group using eqs. (13) and (17), expressed as

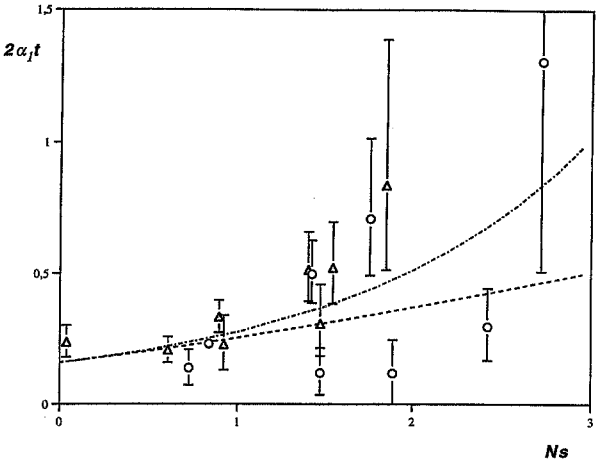

Fig. 1. The correlation between the substitution rate constant to a favored codon estimated from eq. (10) and the selection pressure, $N_e s$, calculated from the codon bias from eq. (18). The *curves* are the theoretical ones from eq. (13) (*dashed*) and eq. (16) (*dash-dot*). Datapoints are from genes at very high (*VH*) expression level (*circles*) and from genes at high (*H*) expression level (*triangles*). The statistical *error bars* were calculated from an assumed binomial sample distribution such that the variance in the observed number (= NP_{Mm}) of differences in a sample of N synonymous sites is $\sigma^2 = NP_{Mm}(1 - P_{Mm})$.

$$2u_1 t = 2\alpha_1 t \frac{1 - B_{VL}/B}{\ln(B/B_{VL})} \tag{18}$$

The results are plotted in Fig. 2; the upper panel shows the mutation rate for A-to-G transitions and the middle panel the mutation rate for T-to-C transitions. Because of the mutational bias, the mutation rates for the C-to-T transitions may be about twice as large as for T-to-C. Although the scatter is large, reflecting the scatter in Fig. 1, the trends are very clear in indicating a monotonic increase in the mutation rate with decreasing expression level.

To further substantiate this systematic variation in the mutation rates, we have also looked at the substitutions between the two major codons ending in T and C in four groups of fourfold degenerate amino acids: threonine, glycine, and the fourfold groups of serine and arginine. The T-to-C substitution rate constant can be calculated with the same formalism as described above if B is the bias for C over T and P_{Mm} is the fraction of C–T differ-

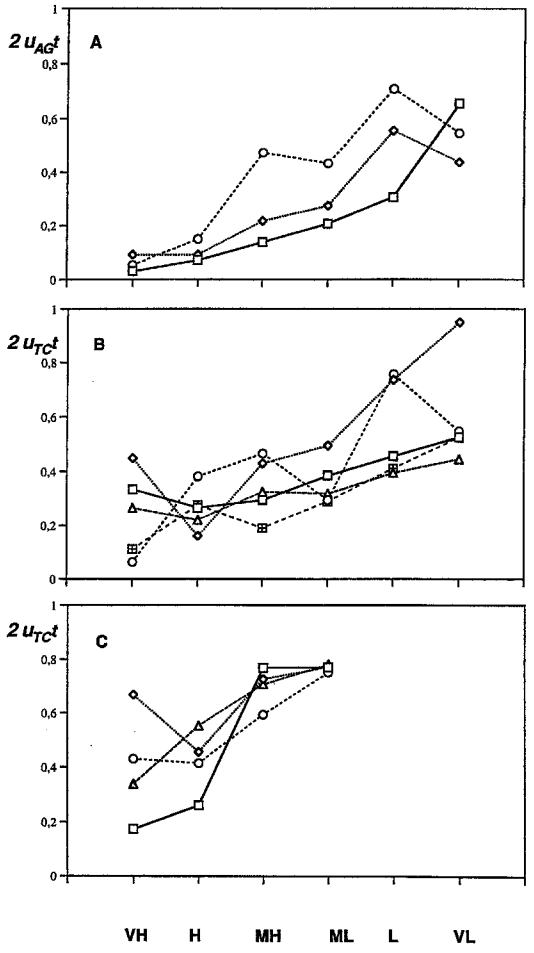

Fig. 2. The calculated mutation rate constants, $2ut$, as function of gene-expression level for the different amino acids as described in the text. **A** is for A-to-G transitions (*squares*—lys, *diamonds*—glu, and *circles*—gln) and **B** is for the T-to-C transitions (*squares*—tyr, *diamonds*—asn, *circles*—his, *triangles*—asp, and *crossed squares*—phe); **A** and **B** are both for the twofold degenerate amino acids. **C** is for the T-to-C transitions in the fourfold degenerate amino acids (*squares*—ser, *diamonds*—thr, *circles*—arg, and *triangles*—gly); for reasons explained in the text, only the four highest-expression-level groups are included in this case.

ences among the pyrimidine-ending codons. The advantage of studying these major–major transitions for the mutation rates is that they are nearly neutral; therefore the correction for selection bias is small and depends less on the assumed form of eq. (13). This calculation does not account for C–T differences that have occurred via two transversional changes; since the fraction of transversional differences is small in all but the two lowest-expression-level groups (L and VL), this can be shown to have a negligible effect on the estimated rates in the other groups. The mutation rate constants for T and C can be calculated from eq. (18) and the result is plotted in the lowest panel of Fig. 2. The resulting tendency for increasing mutation rates at lower expression level in agreement with the results for the twofold degenerate amino acids is obvious. The excluded data points for V and VL would actually strengthen the tendency.

Discussion

Equilibrium Relaxation

The results support the notion that the sequences have evolved since separation with similar selection pressures on synonymous codon choice. Thus, the overall codon distributions have remained at equilibrium and the only rearrangements that have taken place involve changing which codon is at what site. These changes have occurred as small fluctuations around equilibrium when, occasionally, a minor codon has been replaced by a major, or vice versa. The reverse mutation, which keeps the overall codon distribution constant, is more likely to occur at some other site since there are always more other sites. As a consequence of this equilibrium relaxation, the kinetic equations are not required to be valid for large departures from the equilibrium codon distribution.

The agreement with the kinetic model does not prove that the synonymous codon choices are statistical, i.e., that the probability for a certain synonymous codon choice is independent of the position in the gene (beyond the first 100) and independent of codon choices at other positions. We have also looked at the distribution of the distances between successive synonymous changes along the genes (data not shown) and find it to be an exponential distribution. This is consistent with, but does not prove, the idea that the synonymous changes constitute a Poissonian process so that their positions are random. Nonsynonymous changes, however, do not show this characteristic. There remains the possibility that position effects do occur, e.g., as a consequence of constraints in DNA or RNA structure, but they do not appear to be dominant. That the synonymous codon choices are statistical is a working hypothesis that is required for the calculation of the substitution rates; the results presented here are consistent with this hypothesis.

Mutation-Selection Balance

The kinetic equations were developed with the picture of mutation-selection balance in mind. This is apparent, for instance, in eqs. (1)–(8), where the probability of changing a major or a minor codon at a certain site is assumed to be independent of what the distribution is overall. The identifications of the rate constants in eqs. (13–(16) are also based on a mutation-selection balance. In this picture, it is always "best" to use a major codon, and $\alpha_1 > \alpha_2$; however, when there are so many more major codons than minor ones that their numbers exactly compensate the difference in the substitution rates, a balance is reached and the distribution is equilibrated. The effective relaxation rate ($\alpha_1 + \alpha_2$) is dominated by the faster step, α_1, and is therefore expected to be larger than the mutation rate. Furthermore, the relaxation rate is expected to

be faster in genes and/or amino acids at higher selection pressure.

It is useful to consider at least qualitatively also what would happen for stabilizing selection. In this case there would be a strong selection to keep synonymous codons at a fixed and optimal distribution, and all departures from this distribution will be selected against. If the two species had the optimal codon distribution at the time of separation, all subsequent rearrangements would take place via substitutions that were selected against (Kimura 1981, 1983).

The situation is likely to be more complex than either of these two extreme pictures, mutation-selection or stabilizing selection. First of all, if stabilizing selection is to keep the system at the optimum in the way that Kimura (1981) describes it, a very strong selective force would be required to counteract the mutational randomization forces; it does not appear likely that the viability of the organism is so sensitive to small changes in synonymous codon distribution. As a consequence, if indeed an optimum situation exists that is different from the case where all codons are major codons, the system is most likely to be at some distance from it. The result would be a mutation-selection balance.

The selective force on the codon choice depends on the levels of the cognate tRNAs. These levels change with the growth conditions of the cells (Emilsson and Kurland 1990; Emilsson et al. 1993). Thus the resulting force over evolutionary time that determines the codon bias of the organism may represent some average over varying growth conditions. Furthermore, relative tRNA levels could also be changing over evolutionary time, coadapting to changes in the codon distributions (Bulmer 1987). However, since codon bias is approximately the same, though somewhat smaller in *S. typhimurium* than in *E. coli,* there do not appear to have been any major changes in selection pressure and therefore in tRNA levels. The situation seems to be one where tRNA levels are optimal for the codon distribution (Ehrenberg and Kurland 1984), while the codon choices are set by a mutation-selection balance.

Substitution Rates vs. Codon Bias

Previous studies (Sharp and Li 1987; Sharp 1991) of the substitution rates in *E. coli* and *S. typhimurium* showed a strong negative correlation between the average synonymous substitution rate, K_s, and the average codon bias, the codon adaptation index (CAI), in a gene. This result has been quoted as an indication that synonymous codon substitutions are under stabilizing selection (Ikemura 1985). However, the average rate of substitution would have a negative correlation with codon bias even in a model with mutation-selection balance. The appropriate question is whether the rate constant for change to a preferred codon, α_1, correlates positively or negatively with the selection pressure as evidenced in the codon bias. In Kimura's (1981, 1983) model for stabilizing selection α_1 is expected to be negatively correlated with codon bias. By considering separately the correlation between substitution rate constants and codon bias for individual amino acids in different groups of genes we find a largely positive correlation between rate constant and bias, displayed in Fig. 1, as expected from the mutation-selection model, eqs. (13)–(16). However, the correlations disappear if genes at all expression levels are taken together.

When the fraction of synonymous differences, P_{Mm}, is used as a measure of the substitution rate, seemingly contradictory correlations can be found. In fact, rewriting P_{Mm} from eq. (9) as a function of codon bias B by replacing α_1 and $N_e s$ from eqs. (13) and (15), one finds that P_{Mm} increases with increasing bias for low values of B and then decreases for higher values of B. Eyre-Walker and Bulmer (1993) find both negative and positive correlations between the average P_{Mm} and codon adaptation index (CAI): P_{Mm} increases with decreasing CAI as genes with decreasing expression level are considered, while P_{Mm} decreases with decreasing CAI as codons closer to the beginning of the genes are considered. They suggest that the positive correlation is due to some unknown selection pressure that would make the codon choice look more random toward the beginning of the genes. However, in light of the present results we suggest that this positive correlation is totally compatible with the previous model (Bulmer 1988a, 1991) where the selection pressure on codon bias decreases at the beginning of the genes. Furthermore, the increase in P_{Mm} for genes at lower expression (lower CAI) seems to be largely a consequence of the increasing mutation rate constant rather than the decrease in the selection pressure. Thus, the apparently contradictory results of Eyre-Walker and Bulmer (1993) seem to be fully compatible with the kinetic model presented here.

The relationship between substitution rate constants and selection pressure is expected to be a little different if most changes are fixed in a population via hitchhiking (Berg 1995b), eq. (16), rather than through their own genetic drift, eq. (13). However, the difference in rate appears only at sufficiently large selection pressure, and, as is obvious from the error bars in Fig. 1, in this data set the uncertainty is too large to distinguish between eqs. (13) and (16).

The study of the substitution rates cannot separate the parameters N_e and s. Bulmer (1991) has used a model for translational efficiency to estimate the selective values, s, of a major codon in genes at different expression levels. From this he finds that the effective population size for *E. coli* would be about $N_e = 10^5$. This is in qualitative agreement with an estimate based on a mutation-

selection balance for base-pair choices in DNA recognition sites (Berg 1992).

Mutation Rate Constants

By fitting the observed substitution rates to the expected relationships, eq. (13) or (16), it is possible to distinguish between mutation and selection and estimate the underlying mutation rate constants. For the two-codon amino acids studied here this corresponds to a transition rate. The data can be fitted reasonably well (Fig. 1) if the mutation rate constants are different in the genes at different expression levels: The mutation rate for a transition T to C or A to G is $2u_1t = 0.1$–0.2 for genes at very high or high expression level (H or VH), increasing to about $2u_1t = 0.5$ at low or very low expression (L or VL).

According to Table 1, the mutational bias seems to favor T over C in the pyrimidine-ending codons (except asparagine) and A over G in the purine-ending codons (except glutamine). Since all these two-fold degenerate amino acids (except phenylalanine) have an A in the second codon position, this bias for T and A reflects the strong dinucleotide preference in *E. coli* for AT and AA. But there remains a possibility that the mutational bias is different in genes at different expression; this possibility cannot be excluded a priori, in particular as we find the mutation rates to be changing in the different gene groups. Fortunately, the mutational bias enters logarithmically, and small differences would only shift the data points in Fig. 1 small distances along the x-axis.

Increasing expression level seems to have two effects on the synonymous substitution rates: First it increases the selection pressure of synonymous codons, which leads to increased codon bias and also directly influences the substitution rates via the selection coefficient. Second, it decreases the mutation rate constants, either directly or as a secondary effect of the codon bias (Fig. 2). The apparent decrease in the mutation rate constants in genes at high expression level is somewhat unexpected, but not unreasonable. Sharp (1991) detected a decrease in substitution rate around the origin of replication. However, this map-position effect is not likely to be the cause of the decreased mutation rate we observe since the genes in the different groups considered here do not have any systematic distribution around the origin of replication. The most plausible explanation for the apparent variation of the mutation rates is that genes that are transcribed often also are under more rigorous DNA repair. The repair enzymes could find transcribed DNA more accessible and therefore preferentially act on high-expression-level genes; such a coupling between transcription and DNA repair has in fact been identified, at least in some systems (Selby and Sancar 1993). The proposed variation in mutation rates could also, at least partially, explain the observed correlation (Sharp and Li 1987) between synonymous and nonsynonymous substitution rates. Finally, the variation can be considered as a prediction which can be tested with molecular genetic methods.

If *E. coli* and *S. typhimurium* separated around 10^8 years ago (Ochman and Wilson 1987), the average mutation rate constant, u, for transitions would be 10^{-9} per year in genes at high expression level, increasing to about 3×10^{-9} per year in genes at low expression levels. If it is assumed that *E. coli* in "the wild" has about 300 generations per year (Ochman and Wilson 1987), this would correspond to a mutation rate for transitions of about 10^{-11} per generation or less. This is in excellent agreement with recent results from experimental measurements (D. Hughes, personal communication). Since the model cannot distinguish in which organism a change has taken place, these mutation rates actually correspond to the arithmetic mean of their values in the two organisms.

Phylogenetic Distance

The important determinant for the distance between two organisms is the time since separation. To estimate this time based on the observed differences between homologous sequences requires a suitable kinetic model. Synonymous differences can be particularly useful since they may be out of equilibrium between the organisms and still equilibrated within each of them. Usually, synonymous differences are treated as neutral, but in many cases there is a definite selection bias as evidenced in the codon bias. As described above, the formulation of the kinetic model should include the selective bias, if there is any. The molecular clock can be ticking at very different rates for different amino acids and for different genes, even if only silent substitutions are considered.

Acknowledgments. We thank Paul Sharp and Pedro Silva for comments on an earlier version of the manuscript. This work was supported by the Swedish Natural Science Research Council.

References

Berg OG (1992) The evolutionary selection of DNA base pairs in gene-regulatory binding sites. Proc Natl Acad Sci U S A 89:7501–7505

Berg OG (1995a) Kinetics of synonymous codon change for an amino acid of arbitrary degeneracy. J Mol Evol (in press)

Berg OG (1995b) Periodic selection and hitchhiking in a bacterial population. J Theor Biol (in press)

Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. Nature 325:728–730

Bulmer M (1988a) Codon usage and intragenic position. J Theor Biol 133:67–71

Bulmer M (1988b) Are codon usage patterns in unicellular organisms determined by selection mutation balance? J Evol Biol 1:15–26

Bulmer M (1990) The effect of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res 18:2869–2873

Bulmer M (1991) The selection-mutation drift theory of synonymous codon usage. Genetics 129:897–907

Ehrenberg M, Kurland CG (1984) Costs of accuracy determined by a maximal growth rate constant. Q Rev Biophys 17:45–82

Emilsson V, Kurland CG (1990) Growth rate dependence of transfer RNA abundance in *Escherichia coli.* EMBO J 9:4359–4366

Emilsson V, Näslund AK, Kurland CG (1993) Growth-rate dependent accumulation of twelve tRNA species in *Escherichia coli.* J Mol Biol 230:483–491

Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res 21: 4599–4603

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1–21

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc Natl Acad Sci USA 78:5773–5777

Kimura M (1983) The neutral theory of evolution. Cambridge University Press, Cambridge, pp 183–193

Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li WH, Graur D (1991) Fundamentals of molecular evolution. Sinauer Associated, Sunderland, MA p 32

Ochman H, Wilson AC (1987) Evolutionary history of enteric bacteria. In: Neidhardt FC (ed) *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology. ASM Press, Washington, DC, p 1649

Selby CP, Sancar A (1993) Transcription-repair coupling and mutation frequency decline. J Bacteriol 175:7509–7514

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium:* codon usage, map position, and concerted evolution. J Mol Evol 33:23–33

Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592