The Selection-Mutation-Drift Theory of Synonymous Codon Usage

Michael Bulmer[1]

Department of Statistics, Oxford University, Oxford OX1 3TG, England

Manuscript received March 13, 1991
Accepted for publication July 6, 1991

ABSTRACT

It is argued that the bias in synonymous codon usage observed in unicellular organisms is due to a balance between the forces of selection and mutation in a finite population, with greater bias in highly expressed genes reflecting stronger selection for efficiency of translation. A population genetic model is developed taking into account population size and selective differences between synonymous codons. A biochemical model is then developed to predict the magnitude of selective differences between synonymous codons in unicellular organisms in which growth rate (or possibly growth yield) can be equated with fitness. Selection can arise from differences in either the speed or the accuracy of translation. A model for the effect of speed of translation on fitness is considered in detail, a similar model for accuracy more briefly. The model is successful in predicting a difference in the degree of bias at the beginning than in the rest of the gene under some circumstances, as observed in *Escherichia coli*, but grossly overestimates the amount of bias expected. Possible reasons for this discrepancy are discussed.

SYNONYMOUS codons (coding for the same amino acid) are often used with very different frequencies (GRANTHAM *et al.* 1980, 1981; IKEMURA 1985; BULMER 1988a; SHARP 1989; ANDERSSON and KURLAND 1990). The main features of codon usage bias described in these references are as follows. (1) Among a group of synonymous codons recognized by several tRNAs, codons recognized by the most abundant tRNA are used more often than those recognized by rare tRNAs. (2) Among codons recognized by the same tRNA, those making a natural Watson-Crick pair with the anticodon in the wobble position are usually performed, though there are exceptions to this rule (GROSJEAN and FIERS 1982). (3) The bias is more extreme in strongly expressed genes which produce large amounts of protein than in weakly expressed genes. Codon usage bias of this kind has been found in *Drosophila melanogaster* (SHIELDS *et al.* 1988) as well as in many prokaryotes and unicellular eukaryotes but not in mammals.

There is general agreement that the strong bias found in highly expressed genes is due to selection for speed or translational efficiency, but two theories have been put forward to account for codon usage in weakly expressed genes. The first, the *expression-regulation theory*, is that rare codons are used in the latter as a mechanism for keeping their expression low (GROSJEAN and FIERS 1982; KONIGSBERG and GODSON 1983; WALKER, SARASTE and GAY 1984; HIND and BLAKE 1985). The second, the *selection-mutation-drift theory*, is that codon usage patterns result from the balance in a finite population between selection favoring an optimal codon for each amino acid and mutation together with drift allowing the persistence of nonoptimal codons. Selection is likely to be stronger on codons in highly expressed genes since they are used more often, thus generating greater bias in highly than in weakly expressed genes (SHARP and LI 1986a,b; BULMER 1987, 1988a).

Several lines of evidence support the selection-mutation-drift theory rather than the expression-regulation theory of codon usage in weakly expressed genes. (1) The way in which the pattern of codon usage changes with expression level suggests a trend from predominant use of optimal codons in highly expressed genes toward more uniform use of all synonymous codons in weakly expressed genes (SHARP and LI 1986a,b; BULMER 1988a). There is no evidence that weakly expressed genes have a preference for rare codons. (2) The synonymous substitution rate is higher in Enterobacteria in weakly than in highly expressed genes, in accordance with expectation when selection pressure is relaxed (SHARP and LI 1987). (3) Investigation of the effect of neighboring bases (the codon context effect) on the synonymous usage of A *vs.* G or of U *vs.* C shows a similar pattern in complementary sequences in weakly expressed genes in both *Escherichia coli* and yeast (BULMER 1990). This is consistent with an effect of context on mutation rate, but is not consistent with selection acting at the mRNA rather than the DNA level. SHIELDS and SHARP (1987) have made a similar observation in *Bacillus subtilis*. (4) It would seem more efficient to modulate the level of gene expression by changing the strength of the pro-

[1] Current address: Department of Biological Sciences, Rutgers University, Piscataway, New Jersey 08855-1059.

moter or the ribosome binding site. KONIGSBERG and GODSON (1983) discuss a weakly expressed *E. coli* gene sandwiched in the same transcriptional unit between two highly expressed genes, and propose that its weak expression is achieved through the use of non-optimal codons. However, DE BOER and KASTELEIN (1986) point out that it has an inefficient ribosome binding site which is more likely to be the cause of its weak expression. (5) In any case, it seems unlikely that changes in the elongation rate brought about through codon usage will lead to any appreciable change in the rate of protein production (ANDERSSON and KURLAND 1990). The rate of initiation is probably the rate-limiting factor in protein synthesis, so that the elongation rate has little or no direct effect on the rate of protein synthesis. This will be discussed in more detail later.

I shall therefore adopt the selection-mutation-drift theory as a working hypothesis. The purpose of this paper is to evaluate the selective forces likely to act on codon usage in unicellular organisms in which growth rate (or possibly growth yield) can be equated with fitness and to determine whether they are of the right order of magnitude to generate, in conjunction with estimated mutation rates and population size, the observed pattern of codon usage.

The next section describes a simple population genetic model, the following section quantifies from biochemical considerations the selective forces acting on codon usage and the final section confronts the predictions of the model, incorporating these selection coefficients, with the facts. The observed bias in synonymous codon usage is much less than predicted; possible reasons for this discrepancy are considered.

A POPULATION GENETIC MODEL OF SYNONYMOUS CODON USAGE

KIMURA (1981, 1983) considered a model of stabilizing selection on codon usage in which there is an optimal state of highest fitness when the relative frequencies of synonymous codons exactly match those of the cognate tRNAs. The rationale of this model is unclear. It is true that codon usage will alter the relative abundance of cognate charged tRNAs, but it is unlikely that under physiological conditions the tRNA with greatest total abundance (charged, uncharged and bound to mRNA) will also have greatest abundance in free, charged form and will therefore translate fastest regardless of codon usage. This model also fails to take into account differences between codons recognized by the same tRNA, and there is some weakness in the link between its verbal formulation and its mathematical development (LI 1987).

An alternative model will be adopted here in which there is a unique optimal codon for each amino acid, the occurrence of other synonymous codons being due to the combined effects of mutation and genetic drift (LI 1987; BULMER 1987; SHIELDS 1990). Consider first the joint effect of selection and mutation in an effectively infinite population on a two codon family, such as the codons UUU and UUC for Phe. Fix attention on a particular UUY codon at a particular position in a particular gene, and assume that only synonymous changes occur, since nonsynonymous substitutions are comparatively rare. Thus there are two alleles at this codon, represented by the presence of U or C at the third position, which will be labeled B_1 and B_2. For simplicity consider a haploid model, and suppose that individuals carrying B_2 have relative fitness $1 - s$ compared with those carrying B_1, so that s is the selective disadvantage of B_2 compared with B_1, assumed small. Write u for the mutation rate from B_1 to B_2 and v for the mutation rate in the reverse direction. Write p_t and q_t $(=1 - p_t)$ for the relative frequencies of B_1 and B_2 in generation t. The change in the gene frequency in one generation under weak selection and mutation is

$$\Delta p = sp_tq_t + vq_t - up_t. \tag{1}$$

At equilibrium $\Delta p = 0$, so that the equilibrium gene frequency P satisfies the quadratic equation

$$sP(1 - P) + v(1 - P) - uP = 0 \tag{2}$$

whose unique positive root is

$$P = \tfrac{1}{2}\{1 - (u + v)/s + \operatorname{sign}(s)\sqrt{1 - 2(u - v)/s + (u + v)^2/s^2}\}. \tag{3}$$

In a finite population, random sampling introduces stochastic fluctuations in the gene frequency, and we must consider not the equilibrium gene frequency P but the equilibrium distribution of gene frequencies $f(p)$ with Expected value P. A classical result in population genetics (WRIGHT 1931; CROW and KIMURA 1970) is that

$$f(p) \propto e^{Sp}p^{V-1}(1 - p)^{U-1} \tag{4}$$

where

$$\begin{aligned} S &= 2N_es \\ V &= 2N_ev \\ U &= 2N_eu \end{aligned} \tag{5}$$

N_e = effective population size.

If $U + V$ is large, the distribution will be clustered about the deterministic equilibrium in Equation 3. If $U + V$ is small, the population is likely to be at or near one of the boundaries so that the expected gene frequency is the probability of being near 1 rather than 0, given by

$$P = e^SV/(e^SV + U). \tag{6}$$

To prove this result we use the fact (CROW and KI-

MURA 1970, p. 426) that the probability of fixation of a newly arisen mutant with small selective advantage s is approximately

$$\phi(S) = 2S/N(1 - e^{-S}) \quad (7)$$

where N is the actual population size. If the probability of being fixed at 1 is P, the number of new B_2 mutants arising per generation in B_1 populations which are ultimately fixed is $NuP\phi(-S)$; likewise the number of new B_1 mutants arising per generation in B_2 populations which are ultimately fixed is $Nv(1 - P)\phi(S)$. At equilibrium these flux rates are equal, leading to Equation 6.

In conclusion, in a large population ($U + V \gg 1$), we expect to see polymorphism at every codon position, with a fraction of P B_1 codons and $(1 - P)$ B_2 codons, with P given by Equation 3; in a small population ($U + V \ll 1$) we expect to see monomorphism with a fraction P of the relevant positions monomorphic for B_1 and $(1 - P)$ for B_2, with P given by Equation 6.

We have so far only considered selection acting at a single site, whereas we need to consider selection operating simultaneously on a large number of synonymous sites which may be tightly linked (either because they are in the same gene, or because the whole genome is tightly linked in a clonal organism like *E. coli*). If selection acts in a multiplicative way (that is to say, if total fitness is the product of fitnesses at individual sites), no linkage disequilibrium is generated in a deterministic model so that the sites do not interfere with one another and it is legitimate to treat them separately. However, simulations by LI (1987) show that this may not hold for a large number of very tightly linked sites in a finite population. The problem seems to arise from the Hill-Robertson effect (HILL and ROBERTSON 1966; FELSENSTEIN 1974) whereby the existence of background variability between lines in an asexual population increases the variance of the distribution of offspring number and thus decreases the effective population size and the effectiveness of selection. However, this factor will be implicitly allowed for in the present context by estimating the effective population size empirically, so that it seems legitimate to ignore it and to treat the problem with a single site model.

The model can be extended to families of more than two synonymous codons, though the results are more complicated. Suppose that there are K alleles labeled B_1, B_2, . . ., B_K, that the fitness of B_i relative to B_1 is $1 - s_i$ and that the mutation rate from B_i to B_j is u_{ij}. In a finite population write

$$\begin{aligned} S_i &= 2N_e s_i \\ U_{ij} &= 2N_e u_{ij}. \end{aligned} \quad (8)$$

In an infinite population the equilibrium gene frequencies P_i satisfy the set of simultaneous quadratic equations analogous to Equation 3:

$$P_i(\sum_j s_j P_j - s_i) + \sum_j u_{ji} P_j - P_i \sum_j u_{ij} = 0, \quad i = 1, \ldots K. \quad (9)$$

In a finite population in which the U_{ij}s are small enough that there is little polymorphism, the fixation probabilities P_i satisfy the set of simultaneous linear equations analogous to Equation 6:

$$\sum_{j \neq i} (s_i - s_j)(U_{ij} P_i \exp S_i - U_{ji} P_j \exp S_j)/(\exp S_i - \exp S_j) = 0, \; i = 1, \ldots K. \quad (10)$$

SELECTIVE FORCES ACTING ON SYNONYMOUS CODON USAGE

Selection might operate on synonymous codon usage through its effect on mRNA or DNA secondary structure. For example, the secondary structure of mRNA can affect both its rate of degradation (KENNELL 1986) and the rate of initiation of translation (TOMICH *et al.* 1989; GOLD 1988). However, it is not obvious that this type of selection would consistently favor one synonymous codon over another; a suggestion that MS2 phage has biased its codon usage in such a way as to achieve optimal secondary structure (HASEGAWA, YASUNAGA and MIYATA 1979) has proved on reanalysis of the data to be unfounded (BULMER 1989). Selection on secondary structure will therefore be ignored as a major force acting on synonymous codon usage, and attention will be concentrated on factors acting directly on the efficiency of translation through affecting either the speed or the accuracy of translation. Attention will be concentrated on *E. coli* because it has been so intensively studied, but the arguments are sufficiently general that they can be extended to all unicellular organisms (prokaryote or eukaryote). We shall first consider how differences in *speed of translation* between synonymous codons might affect fitness before discussing the possible effects of differences in *accuracy of translation.*

Speed of translation

Codon usage affects speed of translation: It is plausible that codon usage affects speed of translation if a codon recognized by a rare tRNA takes longer to translate than one recognized by a common tRNA. Several lines of evidence suggest that this is the case. (1) Transcription attenuation in, for example, the *leu* operon depends on the fact that the rate of translation of a group of leucine codons in a critical position in the leader decreases as aminoacylated Leu-tRNA^{Leu} becomes rare (LANDICK and YANOFSKY 1987; YANOFSKY 1988). Both the likely mechanism of attenuation and its experimental modulation by changing rare

to synonymous common codons (CARTER, BARTKUS and CALVO 1986; HARMS and UMBARGER 1987; BONEKAMP *et al.* 1985) provide strong evidence of a significant effect of charged tRNA abundance on speed of translation. (2) VARENNE *et al.* (1984) found from studies of incomplete polypeptide chains that their length distribution is not uniform, and they inferred from the pattern of this distribution that the ribosome pauses at sites corresponding to rare codons. (3) SÖRENSEN, KURLAND and PEDERSEN (1989) have measured elongation rate directly *in vivo*, and have found, by inserting a short string of either common or predominantly rare codons into a gene, that rare codons are translated at an average rate of only 2.1 per second compared with 12 per second for the common codons. In a similar experiment, SÖRENSEN and PEDERSEN (1989) have found that strings of GAA, a major codon for Glu, are translated three times faster than strings of GAG, which is recognized by the same tRNA but which is uncommon, particularly in this context (SHPAER 1986). Thus it may be that speed of translation determines codon usage between codons recognized by the same tRNA as well as between those recognized by different tRNAs.

There are two views about how tRNA abundance affects elongation rate. They both assume that activated tRNAs diffuse passively within the cell until a cognate tRNA is sufficiently close to an open A site to be recognized. Under the first model each tRNA within a critical distance of the A site is tested, noncognate ones being rejected, until the first cognate tRNA is encountered and accepted. It is envisaged that the time between the rejection of a noncognate tRNA and the arrival of the next tRNA to be tested is negligible, but that an appreciable time is taken to test each tRNA, during which access to the A site by other tRNAs is blocked (GOUY and GRANTHAM 1980; GOUY and GAUTIER 1982). Under this model it is not the absolute abundance of the cognate tRNA that determines how long a codon takes to be recognized but its relative abundance compared with that of all the other noncognate tRNAs. BILGIN, EHRENBERG and KURLAND (1988) have tested this model by comparing the elongation rate of polyU into polyPhe by the cognate tRNA in an *in vitro* system in the absence and in the presence of a noncognate tRNA. The addition of an excess of noncognate tRNA had a negligible effect on the elongation rate. It is concluded that noncognate tRNAs do not inhibit translation by the cognate tRNA, contrary to the premise of the above model.

Under the second model the time taken to reject noncognate tRNAs is negligible, as suggested by the above experiment, but the time taken until the arrival of the first cognate tRNA in the vicinity of the A site is the rate-limiting factor. This time will be inversely proportional to the absolute abundance of the cognate tRNA. This model presupposes that tRNA abundances are sufficiently low that ribosomes are not saturated by the cognate tRNA. *In vitro* studies of the kinetics of translation support this view (ANDERSSON *et al.* 1986; ANDERSSON and KURLAND 1990). This model can also explain differences in translation rate between codons recognized by the same tRNA, since a codon with a weaker affinity for the anticodon will have a higher K_m and a lower translation rate.

How does speed of translation affect fitness? It is tempting to suppose that a change in the rate of translation (the elongation rate) in a particular mRNA will cause a corresponding change in the rate of protein production by that mRNA; but this is only the case if elongation rather than initiation is the rate-limiting step in protein synthesis. To understand this point, observe that the rate of protein production by messengers of a certain type is equal to the rate of termination on those messengers, which must at equilibrium equal the rate of initiation. A change in the elongation rate will only lead to a change in the rate of protein synthesis if it also leads to a change in the rate of initiation.

If ribosomes are so numerous that a free ribosome is able to bind to the initiation site as soon as it is freed by the movement of the preceeding ribosome, the polysome will be a continuous queue of ribosomes traveling down the messenger in a solid block. In this case elongation is the rate-limiting factor in protein synthesis. An increase in the elongation rate, particularly of the slowest codons, will increase the rate at which the queue of ribosomes travels down the messenger, thus increasing the rate of initiation and hence the rate of protein production. On the other hand, if ribosomes are not saturating so that a free binding site waits an appreciable time before a ribosome binds to it, the polysome will be a free-flowing stream of ribosomes which do not interfere with each other because there are gaps between them. In this case initiation is rate-limiting. Increasing the elongation rate will not directly increase the initiation rate when there is no queue of ribosomes back to the initiation site; the ribosomes will travel faster down the messenger, but there will be fewer of them on the messenger at a time with larger gaps between them. However, this decrease in the polysome size will indirectly increase the rate of initiation on *all* messengers by increasing the pool of free ribosomes, and so will increase slightly the rate of protein production of *all* proteins.

Several lines of evidence suggest that initiation rather than elongation is normally rate-limiting. (1) Since ribosomes form the largest part of the protein translational machinery, it would be inefficient to saturate the system with them. (2) In *E. coli* polysomes

there are about 225 bases per ribosome in a polysome (INGRAHAM, MAALØE and NEIDHARDT 1983, p. 286), while ribosome-protection studies show that the ribosome covers about 30 bases (KOZAK 1983), so that there is an average length of about 195 unoccupied bases between ribosomes. (3) Simulations of models of protein synthesis with realistic parameter values indicate that initiation is rate-limiting (VON HEIJNE, BLOMBERG and LILJENSTRÖM 1987), though HARLEY *et al.* (1981) show that elongation can become rate-limiting if it becomes very slow due to amino acid starvation. (4) Experimental manipulation of codon usage has little effect on protein synthesis except under extreme conditions. For example, ROBINSON *et al.* (1984) inserted four rare arginine codons (AGG) or 4 common arginine codons (CGT) in a cluster into the CAT gene. The insertion made no difference to production of CAT protein at low expression levels, but at very high expression levels in a multicopy plasmid the construct with four AGG codons produced substantially less CAT protein than either the CGT construct or the wild type gene. VARENNE and LAZDUNSKI (1986) have shown by a theoretical calculation that the latter effect is due to the tandem arrangement of the rare codons, which at high expression levels sequester the cognate tRNA in the P-site and so make the translation of these codons rate-limiting. They therefore predicted that the effect would disappear if the AGG codons were scattered throughout the gene rather than clustered; this prediction has been confirmed experimentally (VARENNE *et al.* 1989).

A model for the effect of speed of translation on fitness: LILJENSTRÖM and VON HEIJNE (1987) have proposed a deterministic model of the effect of codon usage on translation time when initiation is rate-limiting, which will be developed further here. Suppose that there are m_i mRNA molecules transcribed from the ith gene and that R_{bi} is the average number of ribosomes bound to an mRNA of this type (the polysome size). If R_{tot} and R_f are the total number of ribosomes and the number of free ribosomes, then

$$R_{tot} = \sum_i m_i R_{bi} + R_f. \tag{11}$$

The time interval, t_{Ii}, between two initiations on a given mRNA of this type is the sum of the time it takes for the newly bound ribosome to move away from the initiation region and of the time it takes for a free ribosome to bind to the messenger once it is unblocked. If we assume that the ribosome blocks further initiation until it has translated L codons and that k_{Ii} is the rate constant for the second process,

$$t_{Ii} = \sum_{j=1}^{L} t_{ij} + (k_{Ii}R_f)^{-1} \tag{12}$$

where t_{ij} is the step time at the jth codon. The total translation time in the absence of queueing is

$$t_{Ti} = \sum_{j=1}^{S_i} t_{ij} \tag{13}$$

if there are altogether S_i codons in the ith gene.

In a steady state

$$R_{bi} = t_{Ti}/t_{Ii} \tag{14}$$

and the rate of synthesis of the ith protein is

$$P_i = m_i/t_{Ii}. \tag{15}$$

Suppose that the step time of the jth codon in the kth mRNA changes because of a synonymous codon change. The change in the relative rate of synthesis of the ith protein (which is a more convenient quantity than the absolute rate of synthesis) is

$$\frac{1}{P_i}\,\partial P_i/\partial t_{kj} = \partial \ln P_i/\partial t_{kj} = -\partial \ln t_{Ii}/\partial t_{kj}. \tag{16}$$

From Equation 12 we can see that two terms are involved. The first term is the direct effect of a change in the step time in the initiation region of the ith mRNA on the production of its own protein because of its effect on the frequency of initiations:

$$\partial \ln P_i/\partial t_{kj}\ (\text{direct}) = -t_{Ii}^{-1} \quad \text{if} \quad k = i,\ j \le L. \tag{17}$$

The negative sign shows that an increase in the step time leads directly to a decrease in protein synthesis, as expected. The second term is the indirect effect of a change in the step time on the production of all proteins through changing the number of free ribosomes (see Equations 11 and 14), which can be found by implicit differentiation:

$$\partial \ln P_i/\partial t_{kj}\ (\text{indirect}) \begin{cases} = (R_{bk} - 1)\alpha_i P_k & j \le L \\ = -\alpha_i P_k & j > L \end{cases} \tag{18}$$

with

$$\alpha_i = 1/[(R_f^2 + \sum_r P_r R_{br} k_{Ir}^{-1}) t_{Ii} k_{Ii}]. \tag{19}$$

The negative sign outside the initiation region ($j > L$) means that an increase in the step time leads indirectly to a decrease in synthesis of all proteins through decreasing the abundance of free ribosomes. On the other hand, the positive sign within the initiation region ($j \le L$) if the polysome size exceeds unity means that an increase in the step time leads indirectly to an increase in synthesis of all proteins through increasing the abundance of free ribosomes; this happens because there are fewer initiations on the kth type of messenger so that its polysome size is reduced.

To interpret these results in terms of relative fitness (w), suppose that α_i has the same value α for all mRNAs so that the indirect effect is the same on the relative production of all proteins. In unicellular or-

ganisms such as *E. coli* and yeast under optimal growth conditions, fitness, growth rate and rate of protein synthesis are nearly synonymous. Thus the change in relative protein synthesis and hence the change in relative fitness resulting from a change in step time outside the initiation region is

$$\partial w/\partial t_{kj} = -\alpha P_k. \tag{20}$$

We shall now try to obtain an approximate estimate of α for *E. coli*. Substituting average values for the gene-specific constants in Equation 19,

$$\alpha = 1/[R_f^2 t_I k_I + R_b t_I P_{\text{tot}}], \tag{21}$$

where P_{tot} is the total rate of protein production. We shall use data for cells in balanced growth at 37° in glucose minimal medium with a mass doubling time of 40 min given by INGRAHAM, MAALØE and NEIDHARDT (1983, pp. 3, 286 and 289) to estimate the parameters in this equation. The number of protein molecules per cell is 2.36×10^6, whence the rate of protein synthesis given a doubling time of 40 min is

$$\begin{aligned} P_{\text{tot}} &= 2.36 \times 10^6 \times \ln 2/(40 \times 60) \\ &= 680 \text{ molecules per sec.} \end{aligned} \tag{22}$$

The number of ribosomes per cell is 18,700 of which about 85% are bound, so that

$$R_f = 0.15 \times 18{,}700 = 2{,}800. \tag{23}$$

The number of messengers per cell is 1380, so that from Equation 15

$$t_I = 1380/680 = 2.0 \text{ sec.} \tag{24}$$

There are about 400 codons per gene and codons are translated at a rate of about 18 per second, so that

$$t_T = 400/18 = 22, \tag{25}$$

and from Equation 14,

$$R_b = 22/2 = 11. \tag{26}$$

The time taken for the ribosome to travel through the initiation region (the first term on the right hand side of Equation 12) may be estimated as about 1 sec by putting $L = 10$ (see earlier) and making some allowance for the fact that translation is slower in this region than in the rest of the gene; hence

$$k_I R_f = 1 \text{ sec.} \tag{27}$$

Substituting these values into Equation 21 we find that

$$\alpha = 0.033/P_{\text{tot}}. \tag{28}$$

From Equation 20 the change in relative fitness resulting from a change in step time outside the initiation region is

$$\partial w/\partial t_{kj} = -0.033\rho \tag{29}$$

where $\rho = P_k/P_{\text{tot}}$ is the relative abundance of the protein (as a proportion of total protein production).

Finally, we calculate the selective disadvantage of a codon outside the initiation region translated by a rare tRNA which slows down the rate of translation sixfold (SÖRENSEN, KURLAND and PEDERSEN 1989) so that the step time is increased by 0.28 sec from 0.05 to 0.33 sec; this would produce a selective disadvantage s of about

$$s = 0.033 \times 0.28\rho = 0.01\rho. \tag{30}$$

It has been assumed in this model that the total number of ribosomes and the tRNA concentrations remain fixed. It would be advantageous for a cell to respond to the introduction of a rare codon by increasing the concentration of the cognate tRNA and the total number of ribosomes to offset the reduction in the rate of protein synthesis. Calculations by O. G. BERG (personal communication) show that, if this response of the protein-synthetic machinery to maintain optimal performance can be made in physiological time by demand-regulation, it will substantially reduce the selection pressure against rare codons. This complication will not, however, be pursued further here.

We turn now to the change in fitness for a codon within the initiation region. This is the sum of two components:

$$\partial w/\partial t_{kj} = (R_{bk} - 1)\alpha P_k - t_{lk}^{-1}\partial w/\partial \ln P_k. \tag{31}$$

The term $\partial w/\partial \ln P_k$ is the change in fitness induced by a change in the production of a single protein. It is likely to be small because changing the activity of one enzyme in a complex metabolic pathway usually has only a small effect on the flux through that pathway (KACSER and BURNS 1973). For example, DEAN, DYKHUIZEN and HARTL (1986) showed that a small increase in β-galactosidase activity in *E. coli* in chemostat cultures with limiting lactose caused an increase in fitness one hundred fold smaller; and they remark that selection under natural conditions must be much less intense than this because *E. coli* inhabits an environment with many alternative sources of carbon and energy.

If initiation is rate-limiting one therefore expects to see different codon usage behavior in the initiation region and in the remainder of the gene in prokaryotes in whom the ribosome binds to an initiation region which covers the beginning of the coding region, as assumed in the above theoretical analysis. However, in eukaryotes the ribosome binds to the cap site at the 5′ end of the mRNA and then migrates down to the AUG start codon (KOZAK 1983); in consequence none of the coding region is within the initiation region as envisaged above so that the effect of translation time on fitness is given by Equation 20 for all codons.

Accuracy of translation

A differential error rate between synonymous codons is the other major selective force likely to affect codon usage. The accuracy of ribosomal translation has been reviewed by BUCKINGHAM and GROSJEAN (1986), Kurland and BLOMBERG (1986) in KIRKWOOD, ROSENBERGER and GALAS (1986). Selection on the ribosome occurs in two stages—an initial discrimination of the ternary complex by the codon at the A site, in which cognate tRNA is more likely to be accepted than noncognate tRNA, followed by proofreading in which noncognate tRNA is more likely to be rejected and released from the A site than cognate tRNA. Write P_j for the probability that initial discrimination leads to acceptance of the jth tRNA species by a particular codon, and Q_j for the probability that this tRNA is rejected at the proofreading stage, with the convention that $j = 1$ denotes the cognate tRNA and $j = 2$ to k denote any noncognate but isoaccepting tRNAs, so that $j > k$ denotes nonisoaccepting tRNAs whose acceptance would lead to an error in translation.

The initial error rate, P_j ($j > 1$), depends on the degree of mismatch between codon and anticodon. For example, the $tRNA^{Leu}$ which recognizes the Leu codons UUA and UUG is accepted initially by the Phe codon UUU about 3% of the time instead of $tRNA^{Phe}$ when the two tRNAs are equally abundant, whereas the $tRNA^{Leu}$ which recognizes the Leu codon CUG is never accepted by UUU (RUUSALA, EHRENBERG and KURLAND 1982). In the above example, the error involved a mispairing in the wobble position, but it can also involve mispairing in the first or second codon positions (BUCKINGHAM and GROSJEAN 1986; MCPHERSON 1988), though an error requiring a mispairing at two positions is probably too rare to be observable. The error rate is also proportional to the relative abundance of the noncognate compared with the cognate tRNA, as expected for a first order reaction (BILGIN, EHRENBERG and KURLAND 1988).

KATO (1990) has suggested that the preference for a codon-anticodon bond of intermediate strength (GROSJEAN and FIERS 1982) is due to avoidance of codons with a high error frequency. For example, in yeast the Gly codon GGU is strongly preferred in highly expressed genes over GGC. These codons are recognized by the abundant $tRNA^{Gly}$ with anticodon GCC (with G in the third position) which is expected to recognize GGC better than GGU. KATO (1990) suggests that this factor, which would lead to quicker translation of GGC by the cognate tRNA, is more than offset by the fact that GGC is much more likely than GGU to be wrongly recognized by the abundant $tRNA^{Asp}$ with anticodon GUC.

In an ideal world, proofreading would always correct a mistake ($Q_j = 1$ for $j > k$). In the real world, this cannot be achieved and there is a tradeoff between maintaining a high probability of correcting a mistake, say $Q_j > 0.99$ for $j > k$, and a low probability of incorrectly rejecting the cognate tRNA, say $Q_1 < 0.05$ (EHRENBERG, KURLAND and BLOMBERG 1986). In consequence the error rate after proofreading is reduced to about 10^{-4}, at the cost of rejecting, say, 5% of cognate tRNAs, which decreases the rate and increases the energetic cost of translation by the same amount.

The average number of trials before a tRNA is finally accepted is

$$N = 1/\sum_j P_j(1 - Q_j). \tag{32}$$

The probability that this tRNA belongs to the jth species is

$$\pi_j = NP_j(1 - Q_j). \tag{33}$$

A perfectly accurate codon would have $N = 1$, $\sum_{j \leq k} \pi_j = 1$. Departures from either of these conditions impose a cost. The excess of N over 1 expresses the cost of proofreading. The deficit of $\sum_{j \leq k} \pi_j$ below 1 measures the frequency with which the wrong amino acid is incorporated. These costs will be discussed in turn.

The cost of proofreading: The term $N - 1$ measures the average number of rejections at proofreading before the final acceptance. Each rejection wastes both time and energy. Since $N - 1$ is likely to be small (less than 0.1), differences between synonymous codons in time wasted by proofreading are likely to be small compared with differences in time taken for the initial discrimination. Only the energy costs will be considered here.

Protein synthesis is very costly in energetic terms; this cost can be expressed as high energy phosphate bonds (~P). Each peptide bond requires 2 ~P for activating aminoacyl-tRNA, 1 ~P for the EF-Tu mediated transfer of the aminoacyl-tRNA to the ribosomal A site and 1 ~P for the EF-G mediated translocation of peptidyl-tRNA to the P site. A total of 4 ~P are thus required for each aminoacyl residue added, of which 3 have already been used if the tRNA is rejected at the proofreading stage. The energy wasted by proofreading by a codon in a gene with relative expression ρ, expressed as a proportion of the total energy used in protein synthesis, is

$$E = 0.75(N - 1)\rho/300 \tag{34}$$

where 300 represents the average number of codons in a gene. (EHRENBERG *et al.* (1990) suggest that 2 ~P are used in the EF-Tu mediated step, in which case the factor 0.75 would be increased to 0.8.)

INGRAHAM, MAALØE and NEIDHARDT (1983) calculate that protein synthesis accounts for about one half of the energy requirement for growth of *E. coli* in minimal medium, and for nearly all the energy re-

quirement for growth in rich medium, so that E should be multiplied by a factor between 0.5 and 1 to express it as a proportion of the total requirement for growth.

A small change in the rate of energy consumption might affect fitness in two ways. First, energy might be the limiting factor for growth rate. It has been suggested by ANDERSEN and VON MEYENBURG (1980) that energy is the limiting factor for growth rate in *E. coli* in aerobic respiration because of the restriction of respiration to the cytoplasmic membrane; this explanation would be limited to aerobic respiration in prokaryotes. Second, in conditions of clonal competition found in unicellular organisms, the major component of fitness might be not the growth rate but the growth yield in a particular medium, which is likely to be strongly correlated with energetic efficiency; if clone A wastes 1% more energy than clone B and if they are growing separately on identical media, then clone A will have 1% fewer individuals than clone B when the carbon content of the medium is exhausted. In this case the selective disadvantage of a rare compared with a common codon can be calculated from Equation 34 if we substitute the difference in the average number of rejections at proofreading between them for $(N-1)$. Taking this difference as 0.4, for example, gives the selective disadvantage as $s = 0.001\rho$. It seems unlikely that the difference is larger than 0.4, so that 0.001ρ is an upper limit on the selective disadvantage.

The cost of translational errors: The term π_j measures the probability that the tRNA finally accepted belongs to the jth species, which will result in the incorporation of an incorrect amino acid if $j > k$. The effect of such an error on fitness will depend on the nature of the error (how similar is the amino acid incorporated to the correct one in chemical terms?), on the nature of the site within the protein (some sites are more highly conserved than others, which suggests that they are more sensitive to change) and on the nature of the protein (some proteins are more highly conserved than others, which suggests that mistakes anywhere within them are less easily tolerated). There is no *a priori* reason to expect the effect of an error on fitness to be proportional to the degree of expression of the protein, but one would expect it to be inversely proportional to the rate of evolution of the protein.

FACING THE FACTS

Highly expressed *E. coli* genes are less biased in their codon usage at the beginning than in the rest of the gene (BULMER 1988b). This is consistent with the model developed here for the effect of speed of translation on fitness, which predicts a difference between the initiation region and the rest of the gene in prokaryotes. This effect was not found in yeast, consistent with the fact that the ribosome binds to the cap site upstream of the start codon in eukaryotes so that no difference between the beginning and the rest of the gene is predicted. However, this phenomenon is not found in all prokaryotes, since SHARP *et al.* (1990) failed to find it in *Bacillus subtilis*. It is possible that accuracy of translation is a more important selective force than speed of translation in *B. subtilis*, affecting fitness either directly or through energetic costs, in which case the initiation region is subject to the same selective force as the rest of the gene.

TABLE 1

The proportion of codons with Y in the third position (P) and the estimated selective advantage ($S = 2N_e s = \ln P/(1-P)$) in two four-codon families in *E. coli*

	Ribosomal protein genes			aa-tRNA synthetase genes		
Codon family	Frequency	P	S	Frequency	P	S
Thr (ACN)	300	0.91	2.3	194	0.79	1.3
Gly (GGN)	467	0.98	3.8	287	0.91	2.3

Let us provisionally assume that speed of translation is the dominant selective force in *E. coli*, so that the selective disadvantage of a rare codon is about 0.01ρ (Equation 21), and investigate whether the frequency of such codons is correctly predicted by the population genetic theory described above. Ribosomal proteins (apart from L7/L12) are each thought to be produced in the same amount, slightly less than 1 per cent of total protein production (INGRAHAM, MAALØE and NEIDHARDT 1983); the selective disadvantage of a ribosomal protein codon outside the initiation region translated by a rare tRNA is thus about 10^{-4}. Aminoacyl-tRNA synthetases are also each produced in similar amounts, about one tenth of that of ribosomal proteins (PEDERSEN *et al.* 1978), so that the comparable disadvantage should be about 10^{-5}. Table 1 shows the usage of Y-ending codons for the four-codon families threonine (ACN) and glycine (GGN) for 45 ribosomal protein genes (excluding L7/L12) and for eight aminoacyl-tRNA synthetase genes, excluding the first 15 codons of each gene. In both families the Y-ending codons (ACY and GGY) are recognized by abundant tRNAs and the R-ending codons (ACR and GGR) by rare tRNAs. These two families were chosen for study because they are unique in showing such a clear-cut pattern of recognition (Table 3 in BULMER 1988a). The quantity $S = 2N_e s$ defined in Equation 5 has been estimated from Equation 6 as

$$S = \ln P/(1-P) \qquad (35)$$

on the assumption that backward and forward mutation rates are equal when only transversions (between

R and Y) are counted as mutations and transitions are ignored.

The results in Table 1 are difficult to reconcile with the theory developed in this paper. Silent site polymorphism is rare in *E. coli* (SAWYER, DYKHUIZEN and HARTL 1987), justifying the use of Equation 6, and its magnitude together with an estimated mutation rate of about 10^{-10} (DRAKE 1974) suggests an effective population size of about 10^9 (see also SELANDER, CAUGANT and WHITTAM 1987). Thus S is predicted to be about 10^5 for ribosomal protein genes and about 10^4 for amino acid (aa)-tRNA synthetase genes, several orders of magnitude larger than the estimated values; in other words no codons at all recognized by rare tRNAs should be observed in a sample of the size of Table 1. To reconcile the estimated values of S with the predicted values would require an effective population size of 10^5 rather than 10^9. It should also be noted that the estimated values of S are only twice as large for ribosomal protein as for aa-tRNA synthetase genes, rather than ten times as large, as predicted if they are proportional to ρ.

Three possible reasons for this discrepancy will now be discussed. First, the calculated selection coefficient may be wrong. An attempt to measure experimentally the protein burden, the cost in terms of fitness of manufacturing a usefuless protein, suggests that it is substantially less than the fraction of waste protein synthesis (KOCH 1983), though the reason for this discrepancy is not understood. The possibility that demand-regulation of the protein-synthetic machinery to maintain optimal efficiency may buffer the impact of codon usage changes on fitness has already been discussed. There is also doubt about the relative contributions to fitness of (1) the effect of speed of translation on growth rate, which might be the dominant factor in optimal growth conditions and would generate a selection coefficient against a rare codon of about 0.01ρ (where ρ is the relative abundance of the protein); (2) the effect of accuracy of translation on energy costs which might be the dominant factor in organisms competing with each other in poor growth conditions and would generate a selection coefficient rather less than 0.001ρ; and (3) the direct effect of accuracy of translation on errors in the protein product, whose effect on fitness it is more difficult to predict. Nevertheless, it is difficult to believe that the selection coefficients calculated above are several orders of magnitude too large. These selection coefficients could be estimated experimentally in defined growth conditions in chemostat experiments by competing strains with a string of either common or rare codons inserted into a gene, as in the experiments of SÖRENSEN, KURLAND and PEDERSEN (1989) on elongation rate discussed above.

A second explanation for the discrepancy is that there may be counter-balancing selection pressures arising from DNA or mRNA secondary structure which in particular positions may tip the balance of selective advantage against the most efficiently translated codon. It has been suggested that a second ribosome binding site exists at the beginning of the coding region in *E. coli* (PETERSEN, STOCKWELL and HILL 1988; SPRENGART, FALSCHER and FUCHS 1990), so that this region might be subject to conflicting selection pressures; this would be an alternative explanation of the lower codon usage bias in this region. Base changes anywhere in the gene may affect mRNA structure in such a way as to affect transcription or mRNA stability, and may thus set up a conflicting selection pressure. Under this explanation rare codons exist because they have some beneficial effect on DNA or mRNA structure; an experimental prediction is that changing them to common codons should increase the translation rate but lower fitness.

The third explanation for the discrepancy is that the population genetic model described above is grossly inadequate because it fails to take into account the genetic structure of clonal organisms, in particular the importance of periodical selection whereby a favorable new mutant periodically sweeps through the population carrying with it associated alleles by hitchhiking (LEVIN 1981; MILKMAN and BRIDGES 1990). Further work on this problem is required.

A similar model has been developed for the evolution of gene-regulatory binding sites (O. G. BERG, unpublished results). Binding sites for a gene-regulatory protein exhibit some variation in DNA sequence which affects their binding strength. The same biological function can be achieved with a strong binding site and a small amount of protein as with a weak binding site and a large amount of protein, but the weak binding site will be at a selective disadvantage because of the associated protein burden. The predicted selection pressure and the expected distribution of binding sites under the joint action of selection, mutation and drift were evaluated. Theoretical predictions are in good agreement with data on *E. coli* binding sites provided that the effective population size is taken to be about 10^4 or 10^5, which seems unreasonably low. This presents the same paradox as that encountered here in developing an evolutionary explanation of synonymous codon usage.

I thank OTTO BERG, TOM KIRKWOOD, CHUCK KURLAND, PAUL SHARP and MONTY SLATKIN for valuable comments and suggestions.

LITERATURE CITED

ANDERSEN, K. B., and K. VON MEYENBURG, 1980 Are growth rates of *Escherichia coli* in batch cultures limited by respiration? J. Bacteriol. **144:** 114–123.

ANDERSSON, D. I., H. W. VAN VERSEVELD, A. H. STOUTHAMER and C. G. KURLAND, 1986 Suboptimal growth with hyperaccurate ribosomes. Arch. Microbiol. **144:** 96–101.

ANDERSSON, S. G. E., and C. G. KURLAND, 1990 Codon preferences in freeliving microorganisms. Microbiol. Rev. **54:** 198–210.

BILGIN, N., M. EHRENBERG and C. G. KURLAND, 1988 Is translation inhibited by noncognate ternary complexes? FEBS Lett. **233:** 95–99.

BONEKAMP, F., H. D. ANDERSEN, T. CHRISTENSEN and K. F. JENSEN, 1985 Codon-defined ribosomal pausing in *Escherichia coli* detected by using the *pyrE* attenuator to probe the coupling between transcription and translation. Nucleic Acids Res. **13:** 4113–4123.

BUCKINGHAM, R. H., and H. GROSJEAN, 1986 The accuracy of mRNA-tRNA recognition, pp. 83–126 in *Accuracy in Molecular Processes*, edited by T. B. L. KIRKWOOD, R. F. ROSENBERGER and D. J. GALAS. Chapman & Hall, London.

BULMER, M., 1987 Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728–730.

BULMER, M., 1988a Are codon usage patterns in unicellular organisms determined by selection mutation balance? J. Evol. Biol. **1:** 15–26.

BULMER, M., 1988b Codon usage and intragenic position. J. Theor. Biol. **133:** 67–71.

BULMER, M., 1989 Codon usage and secondary structure of MS2 phage RNA. Nucleic Acids Res. **17:** 1839–1843.

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

CARTER, P. W., J. M. BARTKUS and J. M. CALVO, 1986 Transcription attenuation in *Salmonella typhimurium*: the significance of rare leucine codons in the *leu* leader. Proc. Natl. Acad. Sci. USA **83:** 8127–8131.

CROW, J. F., and M. KIMURA, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

DE BOER, H. A., and R. A. KASTELEIN, 1986 Biased codon usage: an exploration of its role in optimization of translation, pp. 225–285 in *Maximizing Gene Expression*, edited by W. REZNIKOFF and L. GOLD. Butterworth, Boston.

DEAN, A. M., D. E. DYKHUIZEN and D. L. HARTL, 1986 Fitness as a function of β-galactosidase activity in *Escherichia coli.* Genet. Res. **48:** 1–8.

DRAKE, J. W., 1974 The role of mutation in microbial evolution. Symp. Soc. Gen. Microbiol. **24:** 41–58.

EHRENBERG, M., C. G. KURLAND and C. BLOMBERG, 1986 Kinetic costs of accuracy in translation, pp. 329–361 in *Accuracy in Molecular Processes*, edited by T. B. L. KIRKWOOD, R. F. ROSENBERGER and D. J. GALAS. Chapman & Hall, London.

EHRENBERG, M., A-M. ROJAS, J. WEISER and C. G. KURLAND, 1990 Two EF-Tu's participate in aminoacyl-tRNA binding and peptide bond formation in *E. coli* translation. J. Mol. Biol. **211:** 739–749.

FELSENSTEIN, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

GOLD, L., 1988 Posttranscriptional regulatory mechanisms in *Escherichia coli.* Annu. Rev. Biochem., **57:** 199–234.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7074.

GOUY, M., and R. GRANTHAM, 1980 Polypeptide elongation and tRNA cycling in *Escherichia coli*: a dynamic approach. FEBS Lett. **115:** 151–155.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER and A. PAVE, 1980 Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8:** r49–r62.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE and R. MERCIER, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9:** r43–r79.

GROSJEAN, H., and W. FIERS, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18:** 199–209.

HARLEY, C. B., J. W. POLLARD, C. P. STANNERS and S. GOLDSTEIN, 1981 Model for messenger RNA translation during amino acid starvation applied to the calculation of protein synthetic error rates. J. Biol. Chem. **256:** 10786–10793.

HARMS, E., and H. E. UMBARGER, 1987 Role of codon choice in the leader region of the *ilvGMEDA* operon of *Serratia marcescens.* J. Bacteriol. **169:** 5668–5677.

HASEGAWA, M., T. YASUNAGA and T. MIYATA, 1979 Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res. **7:** 2073–2079.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

HINDS, P. W., and R. D. BLAKE, 1985 Delineation of coding areas in DNA sequences through assignment of codon probabilities. J. Biomol. Struct. Dyn. **3:** 543–549.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13–34.

INGRAHAM, J. L., O. MAALØE and F. C. NEIDHARDT, 1983 *Growth of the Bacterial Cell.* Sinauer, Sunderland, Mass.

KACSER, H., and J. A. BURNS, 1973 The control of flux. Symp. Soc. Exp. Biol. **27:** 65–104.

KATO, M., 1990 Codon discrimination due to presence of abundant non-cognate competitive tRNA. J. Theor. Biol. **142:** 35–39.

KENNELL, D. E., 1986 The instability of messenger RNA in bacteria, pp. 101–142, in *Maximizing Gene Expression*, edited by W. REZNIKOFF and L. GOLD. Butterworth, Boston.

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 Molecular Evolution.* Cambridge University Press, Cambridge.

KIRKWOOD, T. B. L., R. F. ROSENBERGER and D. J. GALAS, 1986 *Accuracy in Molecular Processes.* Chapman & Hall, London.

KOCH, A. L., 1983 The protein burden of *lac* operon products. J. Mol. Evol. **19:** 455–462.

KONIGSBERG, W. J. N., and G. N. GODSON, 1983 Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli.* Proc. Natl. Acad. Sci. USA **80:** 687–691.

KOZAK, M., 1983 Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev. **47:** 1–43.

KURLAND, C. G., and J. A. GALLANT, 1986 The secret life of the ribosome, pp. 127–157 in *Accuracy in Molecular Processes*, edited by T. B. L. KIRKWOOD, R. F. ROSENBERGER and D. J. GALAS. Chapman & Hall, London.

LANDICK, R., and C. YANOFSKY, 1987 Transcription attenuation, pp. 1276–1307 in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, edited by F. C. NEIDHARDT. ASM Press, Washington, D.C.

LEVIN, B. R., 1981 Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics **99:** 1–23.

LI, W-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

LILJENSTRÖM, H., and G. VON HEIJNE, 1987 Translation rate modification by preferential codon usage: intragenic position effects. J. Theor. Biol. **124:** 43–55.

MCPHERSON, D. T., 1988 Codon preference reflects mistranslational constraints: a proposal. Nucleic Acids Res. **16:** 4111–4120.

MILKMAN, R., and M. M. BRIDGES, 1990 Molecular evolution of

the *Escherichia coli* chromosome. III Clonal frames. Genetics **126:** 505–517.

PEDERSEN, S., P. L. BLOCH, S. REEH and F. C. NEIDHARDT, 1978 Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. Cell **14:** 179–190.

PETERSEN, G. B., P. A. STOCKWELL and D. F. HILL, 1988 Messenger RNA recognition in *Escherichia coli*: a possible second site of interaction with 16S ribosomal RNA. EMBO J. **7:** 3957–3962.

ROBINSON, M., R. LILLEY, S. LITTLE, J. S. EMTAGE, G. YARRANTON, P. STEPHENS, A. MILLICAN, M. EATON and G. HUMPHREYS, 1984 Codon usage can affect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Res. **12:** 6663–6671.

RUUSALA, T., M. EHRENBERG and C. G. KURLAND, 1982 Is there proofreading during polypeptide synthesis? EMBO J. **1:** 741–745.

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

SELANDER, R. K., D. A. CAUGANT, and T. S. WHITTAM, 1987 Genetic structure and variation in natural populations of *Escherichia coli*, pp. 1625–1648 in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, edited by F. C. NEIDHARDT. ASM Press, Washington, D.C.

SHARP, P. M., 1989 Evolution at 'silent' sites in DNA, pp. 23–32 in *Evolution and animal breeding: Reviews on Molecular and Quantitative Approaches in Honour of Alan Robertson*, edited by W. G. HILL, and T. F. C. MACKAY. CAB International, Wallingford, U.K.

SHARP, P. M., and W-H. LI, 1986a An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

SHARP, P. M., and W-H. LI, 1986b Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. Nucleic Acids Res. **14:** 7737–7749.

SHARP, P. M., and W-H. LI, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 22–230.

SHARP, P. M., D. G. HIGGINS, D. C. SHIELDS, K. M. DEVINE and J. A. HOCH, 1990 *Bacillus subtilis* gene sequences, pp. 89–98 in *Genetics and Biotechnology of Bacilli*, vol. 3, edited by M. M. ZUKOSKI, A. T. GANESAN and J. A. HOCH. Academic Press, San Diego.

SHIELDS, D. C., 1990 Switches in species-specific codon preferences: the influence of mutation biases. J. Mol. Evol. **31:** 71–80.

SHIELDS, D. C., and P. M. SHARP, 1987 Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutation biases. Nucleic Acids Res. **15:** 8023–8040.

SHIELDS, D. C., P. M. SHARP, D. C. HIGGINS and F. WRIGHT, 1988 Silent sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

SHPAER, E. G., 1986 Constraints on codon context in *Escherichia coli* genes: their possible role in modulating the efficiency of translation. J. Mol. Biol. **188:** 555–564.

SÖRENSEN, M. A., C. G. KURLAND, and S. PEDERSEN, 1989 Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. **207:** 365–377.

SÖRENSEN, M. A., and S. PEDERSEN, 1989 Abstracts, Thirteenth International Transfer RNA Meeting, Vancouver. (Quoted in ANDERSSON and KURLAND, 1990).

SPRENGART, M. L., H. P. FALSCHER and E. FUCHS, 1990 The initiation of translation in *E. coli*: apparent base pairing between the 16srRNA and downstream sequences of the mRNA. Nucleic Acids Res. **18:** 1719–1723.

TOMICH, C-S., E. R. OLSON, M. K. OLSEN, P. S. KAYTES, S. K. ROCKENBACH and N. T. HATZENBUHLER, 1989 Effect of nucleotide sequences directly downstream from the AUG on the expression of bovine somatotropin in *E. coli*. Nucleic Acids Res. **17:** 3179–3197.

VARENNE, S., and C. LAZDUNSKI, 1986 Effect of distribution of unfavourable codons on the maximum rate of gene expression by an heterologous organism. J. Theor. Biol. **120:** 99–110.

VARENNE, S., J. BUC, R. LLOUBES and C. LAZDUNSKI, 1984 Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. **180:** 549–576.

VARENNE, S., D. BATY, VERHEIJ., D. SHIRE and C. LAZDUNSKI, 1989 The maximum rate of gene expression is dependent on the downstream context of unfavourable codons. Biochimie **71:** 1221–1229.

VON HEIJNE, G., C. BLOMBERG and H. LILJENSTRÖM, 1987 Theoretical modelling of protein synthesis. J. Theor. Biol. **125:** 1–14.

YANOFSKY, C., 1988 Transcription attenuation. J. Biol. Chem. **263:** 609–612.

WALKER, J. E., M. SARASTE and N. J. GAY, 1984 The *unc* operon, nucleotide sequence, regulation and structure of ATP-synthase. Biochim. Biophys. Acta **768:** 164–200.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Communicating editor: W-H. LI