J Mol Evol (2001) 53:225–236
DOI: 10.1007/s002390010212

JOURNAL OF MOLECULAR EVOLUTION

Why Are Translationally Sub-Optimal Synonymous Codons Used in *Escherichia coli*?

Nick G.C. Smith, Adam Eyre-Walker

Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton, BN1 9QG, UK

Received: 18 July 2000 / Accepted: 17 April 2001

Abstract. Natural selection favors certain synonymous codons which aid translation in *Escherichia coli*, yet codons not favored by translational selection persist. We use the frequency distributions of synonymous polymorphisms to test three hypotheses for the existence of translationally sub-optimal codons: (1) selection is a relatively weak force, so there is a balance between mutation, selection, and drift; (2) at some sites there is no selection on codon usage, so some synonymous sites are unaffected by translational selection; and (3) translationally sub-optimal codons are favored by alternative selection pressures at certain synonymous sites. We find that when all the data is considered, model 1 is supported and both models 2 and 3 are rejected as sole explanations for the existence of translationally sub-optimal codons. However, we find evidence in favor of both models 2 and 3 when the data is partitioned between groups of amino acids and between regions of the genes. Thus, all three mechanisms appear to contribute to the existence of translationally sub-optimal codons in *E. coli.*

Key words: Codon bias — Translational selection — Mutation selection drift — *E. coli* — Polymorphism

Introduction

There can be little doubt that synonymous codon bias is the consequence of natural selection in *Escherichia coli*: there is preferential use of codons which match the commonest tRNA or bind the tRNA with optimal base pairing (Ikemura 1985), and the level of synonymous codon bias is strongly correlated to gene expression level (Gouy and Gautier 1982; Ikemura 1985). These two observations suggest that it is some factor during translation which exerts selection upon synonymous codon use in *E. coli*.

However, we still do not know what the precise basis of the selection might be. It has been suggested that selection might be acting on the rate of elongation, the cost of proof-reading, or the accuracy of translation (Bulmer 1991), but it has proved difficult to differentiate between them (Akashi and Eyre-Walker 1998). It is also unclear whether selection for translational efficiency is the only selective force acting upon synonymous codon use; it has been suggested there may be alternative conflicting selection pressures acting upon synonymous codon use, such as selection for the regulation of gene expression, or selection upon mRNA and DNA secondary structure (Eyre-Walker 1996; Eyre-Walker and Bulmer 1995; Hartl et al. 1994).

These issues relating to the nature of selection impinge upon the problem addressed in this paper: why are sub-optimal codons found given that some codons are evidently optimal for translation? Translationally sub-optimal codons are always found, even in the most highly expressed genes (Eyre-Walker 1996). Throughout this paper we will use the terms optimal and sub-optimal to refer to translational selection alone. Thus a translationally sub-optimal codon may be preferred at some site for some other selective reason, for example, it may encode part of a ribosomal binding site, but it would still be referred to as a sub-optimal codon.

Correspondence to: N.G.C. Smith; *email:* n.g.c.smith@sussex.ac.uk

There are at least four explanations for why sub-optimal codons persist. First, selection might favor optimal codons, but be so weak that sub-optimal codons can be fixed by genetic drift; there would then be a balance between mutation, selection, and genetic drift. Second, selection might consistently favor the optimal codon, but vary between sites so that selection is ineffective at some sites, while very effective at others; synonymous codon bias would be caused by the sites at which selection was strong, while sub-optimal codons would persist at the sites where selection was effectively neutral. This is the situation we might expect if selection acts upon translational accuracy, since errors at some codons are more costly than at others. Third, selection might vary between sites, but not always favor the optimal codon, because of conflicting selection pressures, for example, a codon might form part of the ribosome binding site (Eyre-Walker 1996; Eyre-Walker and Bulmer 1993). Fourth, sub-optimal codons might exist because the system is not at equilibrium; this could be because the selection pressures on synonymous codon use have changed, as we see in *Drosophila melanogaster* (Akashi 1996), or because of some other process such as amino acid substitution which can generate sub-optimal codons if different amino acids have optimal codons which differ at their degenerate sites.

These hypotheses are not mutually exclusive. However, we can characterize the main reasons for the existence of sub-optimal codons in sequences as follows; sub-optimal codons may exist because there is a balance between mutation, selection, and genetic drift; there is no selection on synonymous codon use at some codons; there is selection favoring the sub-optimal codon at some sites; and the system is not at equilibrium. We call these the *MSD* (Mutation-Selection-Drift), *neutral*, *conflict*, and *non-equilibrium* hypotheses, respectively.

In this paper we develop methods to discriminate between the MSD, neutral, and conflict models using the frequency distribution of optimal codons in samples taken from a single population. The comparison of frequency distributions has proved a useful tool for the elucidation of selection in different species and at different classes of sites (Akashi 1994; Akashi 1995; Akashi and Schaeffer 1997; Hartl et al. 1994; Sawyer and Hartl 1992; Sawyer et al. 1987). However, such studies have employed models of irreversible mutation, which means that mutations must be polarized. In other words, the direction of mutation must be accounted for, either by using an outgroup sequence to infer the direction of mutation (synonymous mutations from a sub-optimal codon to an optimal codon are defined as preferred, and mutations in the opposite direction are unpreferred, see Akashi 1995), or by consolidating frequency data to combine alternative directions of mutation (in a sample of n sequences, polymorphisms segregating in r sequences are consolidated with polymorphisms segregating in n-r sequences, see Hartl et al. 1994). In contrast, we have used a model of reversible mutation, which eliminates the requirement for either outgroup sequence data or the consolidation of frequency data. Our data consists of multiple sequences of several *E. coli* genes, from which we derive distributions of the frequencies of optimal codons.

We then use the frequency distribution data to discriminate between the MSD, neutral and conflict hypotheses for the existence of suboptimal codons. The MSD and neutral hypotheses make simple testable predictions about the pattern of polymorphism we expect to see in a population. If we just consider sites at which we have optimal and sub-optimal codons segregating in the population, then we expect to see the optimal codon at higher frequency under the MSD hypothesis, since optimal codons are mildly advantageous, and sub-optimal codons are mildly deleterious. In contrast, under the neutral hypothesis we expect the average frequency of optimal and sub-optimal codons at segregating sites to be similar, since the sites where selection is strong will contribute little to polymorphism, leaving the neutral sites to contribute most of the variation. The conflict hypothesis makes less clear-cut predictions: whether translationally optimal or sub-optimal codons are found at higher frequencies at segregating sites depends on the relative strengths of the conflicting selection pressures. We do not test the non-equilibrium hypothesis (but see Discussion).

A Population Genetics Model of Synonymous Codon Evolution

Following Li (1987) and Bulmer (1991) let us imagine that each site has two alleles, *A1* and *A2*, where *A1* is preferred for translation and has a selective advantage, s, over *A2*. Let the mutation rate from *A1* to *A2* be u, and the mutation rate in the opposite direction be v. Under this model Wright (1949) showed that the equilibrium distribution of gene frequencies at a single site in a haploid population is $F(x)$ where

$$F(x) = CE^{Sx}x^{(V-1)}(1-x)^{(U-1)} \qquad (1)$$

where $S = 2N_e s$, $U = 2N_e u$, $V = 2N_e v$, C is a constant which normalizes the function so the integral sums to one, and x is the frequency of the *A1* allele. Under the infinite sites assumption, (i.e. U and V tend to zero) and the assumption of unbiased mutation ($U = V$), the equation can be simplified (McVean and Vieira 1999).

$$F(x) = CE^{Sx}x^{-1}(1-x)^{-1} \qquad (2)$$

If we assume that sites evolve independently, if there is free recombination and no epistasis between sites, then

by the principle of ergodicity (Ewens 1979) we can use the expected pattern at a single site to predict the pattern at many sites (McVean and Vieira 1999). The expected proportion of segregating sites with *i* optimal alleles segregating in a sample of *n* sequences, $G(i)$, is found using a binomial probability model (Hartl et al. 1994), assuming random sampling from the population.

$$G(i) = C \int_{x=1/Ne}^{1-1/Ne} \frac{n!}{i!(n-i)!} x^i (1-x)^{n-i} F(x) dx \quad (3)$$

C is a constant which ensures that $\Sigma_{i=1}^{n-1} G(i) = 1$. The value of $G(i)$ is highly insensitive to changes in N_e as long as N_e is large, and in this study we have used $N_e = 10^6$ throughout. The expected frequency of preferred alleles at polymorphic sites is given by K.

$$K = \sum_{i=1}^{n-1} \frac{G(i)i}{n} \quad (4)$$

The MSD Model

Under the simplest MSD model the strength of selection in favor of the optimal codon, S, is same for all sites. As can be seen from Fig. 1, the frequency of the optimal codon at segregating sites, K, increases as S increases. More precisely, changes in S affect the shape of the distribution of segregating sites, with high positive S yielding a distribution skewed towards optimal alleles segregating at high frequencies (Fig. 2).

The Neutral Model

The neutral model assumes that all sub-optimal codons are at neutral sites, and that translational selection at other sites is sufficiently strong that such sites make no contribution to polymorphism. Thus, the predictions of the neutral model are equivalent to a MSD model for which $S = 0$ for all segregating polymorphisms, which means that $K = 0.5$ (Fig. 1) and that the shape of the distribution of segregating sites is symmetrical (Fig. 2).

The Conflict Model

There are many different conflict models including ones which generate sub-optimal codons because selection is weak or non-existent at some sites. However, the simplest conflict model is one which does not require genetic drift and mutation for the persistence of sub-optimal codons; sub-optimal codons exist because there is strong selection favoring them. A simple conflicting selection pressures model is as follows; let us imagine that one half of the sites are subject to conflicting selection pressures, and that the conflicting selection pressure favors the optimal codon, *A1*, in half the cases, and the sub-optimal codon, *A2*, in the other half. Let S_t be the translational selection in favor of the optimal codon, and S_c be the strength of the conflicting selection pressure ($S_t >> 1$ and $S_c >> 1$ because selection is strong).

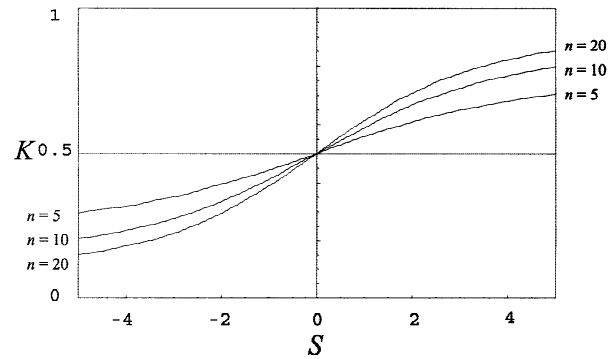

Fig. 1. The expected relationship under the MSD model between S, the product of effective population size and the selection coefficient affecting codon usage, and K, the frequency of the selected allele at polymorphic sites (derived from Equations 2–4). The expected relationship is shown for three different samples sizes, n. Note that $K = ½$ when $S = 0$, and that K covaries with S.

Thus, we have three classes of sites, for each of which we apply the MSD model with different selection coefficients: (1) translational selection only, $S = S_t$, (2) translational and conflicting selection working in the same direction, $S = S_t + S_c$, (3) translational and conflicting selection working in different directions, $S = S_t - S_c$. We know from our model that these selection pressures apply at ½, ¼, and ¼ of all sites respectively, but we also need to account for the differing contributions of the different classes to polymorphism in order to predict $G(i)$ for the conflict model: the class of site at which selection is weakest will make the greatest contribution to polymorphism (see Appendix for details).

If some sites are to be fixed for sub-optimal codons then S_c must be greater than S_t, assuming that the alternative selection pressures are unlikely to balance each other exactly. If $S_c >> 2S_t$, then the sites under the weakest selection are those affected by translational selection only ($S_c + S_t > S_c - S_t > S_t$), and so the greatest contribution to polymorphism is by selection in favor of optimal codons, and thus, we find most of the polymorphisms segregating at $i = n - 1$ (conflict model 1 in Fig. 2: $S_t = 200$, $S_c = 500$). If $S_c << 2S_t$, then the sites under the weakest selection are those at which the alternative selection pressures are conflicting ($S_c + S_t > S_t > S_c - S_t$), and so the greatest contribution to polymorphism is by selection in favor of non-optimal codons, and thus, we find most of the polymorphisms segregating at $i = 1$ (conflict model 2 in Fig. 2: $S_t = 200$, $S_c = 250$). Therefore, we predict that most polymorphisms will be singletons if the strong selection conflict model is correct, in contrast to the MSD model which suggests a greater distribution of polymorphism (see Fig. 2).

Materials and Methods

We have defined optimal codons as those codons which increase in frequency between high and low expression level genes (see Table 1), using the codon usage data given by Sharp et al. (1992). This classification assumes that translational selection is the dominant form of selection upon synonymous codon use. We classified 18 amino acids containing a total of 59 codons (64 codons minus TAA, TAG, TGA, TGG, and ATG), calculating the relative synonymous codon usage (the observed number of codons divided by the number expected if codon usage is random) in the high and low expression level genes, and also the ratio of the two quantities. If the high/low ratio is greater than one the codon increases in frequency with expression level and is defined as an optimal codon, and if the ratio is less than one the codon decreases in frequency with expression level and is defined as sub-optimal. For each amino acid we calculated the optimal/sub-optimal ratio, defined as the mean high/low ratio for optimal codons divided by the mean high/low ratio for sub-optimal codons. By definition, the optimal/sub-optimal ratio must be greater than one, and the greater the ratio the greater the effect of expression levels on codon usage. We have ranked the amino acids according to their optimal/sub-optimal ratios, and have classified the amino acids into high and low optimal/sub-optimal groups of six and twelve amino acids, respectively. The amino acids were divided into the two groups on the basis of the discontinuity in the optimal/sub-optimal ratio between serine and isoleucine.

We compiled multiple sequences for 11 *E. coli* genes from the literature: *celC* (Hall and Sharp 1992), *crr* (Hall and Sharp 1992), *gapA* (Guttman and Dykhuizen 1994b; Nelson et al. 1991), *gnd* (Bisercic et al. 1991; Dykhuizen and Green 1991), *gutB* (Hall and Sharp 1992), *mdhA* (Boyd et al. 1994; Vogel et al. 1987), *pabB* (Guttman and Dykhuizen 1994b), *phoA* (Dubose et al. 1988), *putP* (Nelson and Selander 1992), *sppA* (Guttman and Dykhuizen 1994a), and *zwf* (Guttman and Dykhuizen 1994a).

For each gene we considered all the synonymous sites at which an optimal and a sub-optimal codon were segregating, or at which a single optimal or sub-optimal codon were fixed. From these data we obtained the observed frequency spectrum for each gene, $N(i)$. For a sample of n sequences, $N(i)$ is defined from $i = 0$ to $i = n$ as the number of sites at which an optimal codon is found at i sequences. We used the frequency spectra to calculate a number of statistics: π, the nucleotide diversity at synonymous sites, F, the proportion of sites fixed for the optimal codon, and K, the mean frequency of optimal codons at segregating sites (see Table 2). With one exception, *gnd* ($\pi = 0.188$), the nucleotide diversities are low, indicating that the infinite sites assumption is met (excluding *gnd*, mean $\pi = 0.036$). The gene *gnd* is not considered in the following analyses.

In some analyses we have used a *combined* dataset. This was generated by choosing eight samples at random for those genes with more than eight samples. Then the data from all the genes was summed to give the combined spectrum, $N(i)$ from $i = 0$ to $i = 8$: {938, 27, 22, 11, 21, 20, 31, 48, 1553} (see Table 3).

Results

The Average Frequency of the Translationally Optimal Codon at Segregating Sites

Table 2 shows that sub-optimal codons are present at appreciable frequencies in all the genes we have studied; on average $F = 0.62$. This is the result we seek to explain: why is F less than one? Let us first consider the average frequency of the optimal codon at segregating sites, K. K is greater than ½ in seven out of ten genes, and overall K is significantly greater than 0.5 (mean $K = 0.57$, Wilcoxon signed rank test, $p < 0.001$), as expected under the MSD model and some conflict models (for example conflict model 1 in Fig. 2). However, the results suggest that the neutral model cannot by itself explain why sub-optimal codons are found in *E. coli* genes and its seems unlikely that the conflict model can explain the data either, since singletons do not constitute the majority of polymorphisms.

Maximum Likelihood Estimates of S *Under the MSD Model*

The polymorphism data appear to be inconsistent with both the neutral and conflict models, but are the data really consistent with the MSD model? To investigate this question we ask whether the MSD model provides a good fit to the data. The likelihood, L, of the observed polymorphism data, $N(i)$, given the predicted distribution of polymorphism $G(i)$ (Equation 3), is calculated as a multinominal distribution.

$$L = \frac{\left(\sum_{i=1}^{n-1} N(i)\right)!}{\prod_{i=1}^{n-1} N(i)!} \prod_{i=1}^{n-1} G(i)^{N(i)} \tag{5}$$

$$LogL = Log\left(\left(\sum_{i=1}^{n-1} N(i)\right)!\right) + \sum_{i=1}^{n-1} (N(i)Log(G(i)) - Log(N(i)!)$$

Under the MSD model, $G(i)$ is determined for a given value of n by the coefficient of selection S. So we can estimate S by finding the value which maximizes the likelihood of observing the polymorphism data. Confidence intervals can be determined on the basis of double the difference in $LogL$ being approximately χ^2 distributed: 95% confidence intervals are given by the range of values of S with $LogL$ less than 1.92 below the maximum value (since $\chi^2(0.05) = 3.84$). However, the lack of free recombination in *E. coli* is likely to cause us to underestimate the size of the confidence intervals because we have not taken into account the variance associated with the evolutionary process.

The values of S estimated under the MSD model are given in Table 2 (S_{ML} values). Seven genes yielded an estimate of S greater than zero, although only *putP*, *mdhA*, and *phoA* gave estimates of S significantly greater than zero. However, an estimate of the average selection strength, found by assuming that S is the same across all genes, is significantly greater than zero ($S = 0.80$, $\Delta LogL = 5.20$, $p = 0.0013$).

The data suggest a tendency for the strength of selec-

Table 1. Relative synonymous codon usage

Amino Acid	Codon	Codon Class	Expression Level[1]			Optimal/Sub-Optimal Ratio	Optimal/Sub-Optimal Rank, Class
			Low	High	Ratio		
Lys	AAA	optimal	1.42	1.61	1.13	1.69	1, low
	AAG	sub-optimal	0.58	0.39	0.67		
Val	GTT	optimal	1.43	1.85	1.29	1.69	2, low
	GTA	optimal	0.91	0.96	1.05		
	GTC	sub-optimal	0.75	0.36	0.48		
	GTG	sub-optimal	0.91	0.83	0.91		
Ala	GCT	optimal	1.05	1.39	1.32	2.05	3, low
	GCG	optimal	0.81	1.13	1.40		
	GCC	sub-optimal	0.83	0.46	0.55		
	GCA	sub-optimal	1.32	1.02	0.77		
Glu	GAA	optimal	1.29	1.58	1.22	2.07	4, low
	GAG	sub-optimal	0.71	0.42	0.59		
Cys	TGC	optimal	0.80	1.21	1.51	2.30	5, low
	TGT	sub-optimal	1.20	0.79	0.66		
Gln	CAG	optimal	1.09	1.68	1.54	4.38	6, low
	CAA	sub-optimal	0.91	0.32	0.35		
Asp	GAC	optimal	0.53	1.23	2.32	4.43	7, low
	GAT	sub-optimal	1.47	0.77	0.52		
Thr	ACT	optimal	1.11	1.37	1.23	7.18	8, low
	ACC	optimal	0.87	2.22	2.55		
	ACA	sub-optimal	1.36	0.12	0.09		
	ACG	sub-optimal	0.66	0.29	0.44		
His	CAC	optimal	0.65	1.57	2.42	7.58	9, low
	CAT	sub-optimal	1.35	0.43	0.32		
Tyr	TAC	optimal	0.50	1.44	2.88	7.71	10, low
	TAT	sub-optimal	1.50	0.56	0.37		
Phe	TTC	optimal	0.59	1.57	2.66	8.73	11, low
	TTT	sub-optimal	1.41	0.43	0.30		
Ser	TCT	optimal	1.17	2.32	1.98	10.49	12, low
	TCC	optimal	0.62	1.86	3.00		
	AGC	optimal	0.95	1.28	1.35		
	TCA	sub-optimal	1.49	0.17	0.11		
	TCG	sub-optimal	0.61	0.23	0.38		
	AGT	sub-optimal	1.16	0.13	0.11		
Ile	ATC	optimal	0.64	2.30	3.59	14.69	13, high
	ATT	sub-optimal	1.41	0.69	0.49		
	ATA	sub-optimal	0.95	0.00	0.00		
Leu	CTG	optimal	1.34	5.08	3.79	15.34	14, high
	TTA	sub-optimal	1.63	0.10	0.06		
	TTG	sub-optimal	0.92	0.16	0.17		
	CTT	sub-optimal	1.09	0.25	0.23		
	CTC	sub-optimal	0.52	0.38	0.73		
	CTA	sub-optimal	0.50	0.02	0.04		
Pro	CCG	optimal	0.72	3.06	4.25	16.31	15, high
	CCT	sub-optimal	1.10	0.38	0.35		
	CCC	sub-optimal	0.86	0.03	0.03		
	CCA	sub-optimal	1.32	0.53	0.40		
Asn	AAC	optimal	0.59	1.78	3.02	19.34	16, high
	AAT	sub-optimal	1.41	0.22	0.16		
Gly	GGT	optimal	1.16	2.19	1.89	31.07	17, high
	GGC	optimal	0.82	1.70	2.07		
	GGA	sub-optimal	1.16	0.04	0.03		
	GGG	sub-optimal	0.86	0.08	0.09		
Arg	CGT	optimal	1.42	4.11	2.89	104.47	18, high
	CGC	optimal	1.06	1.81	1.71		
	CGA	sub-optimal	0.79	0.03	0.04		
	CGG	sub-optimal	0.76	0.02	0.03		
	AGA	sub-optimal	1.26	0.03	0.02		
	AGG	sub-optimal	0.70	0.00	0.00		

[1] Data based on Sharp et al. (1992).

Table 2. Synonymous codon usage statistics

Gene	π	*F*	*K*	*n*	S_{ML}	S_K
celC	0.032	0.49	0.59	11	0.96	0.98
crr	0.034	0.74	0.67	12	1.83	1.85
gapA	0.009	0.91	0.68	18	1.73	1.75
gutB	0.024	0.57	0.49	11	−0.10	−0.10
mdhA	0.028	0.72	0.73	16	2.36	2.41
pabB	0.037	0.52	0.54	12	0.46	0.45
phoA	0.050	0.53	0.60	8	1.28	1.29
putP	0.070	0.53	0.59	8	1.11	1.09
sppA	0.032	0.61	0.43	12	−0.69	−0.69
zwf	0.036	0.62	0.48	11	−0.20	−0.20

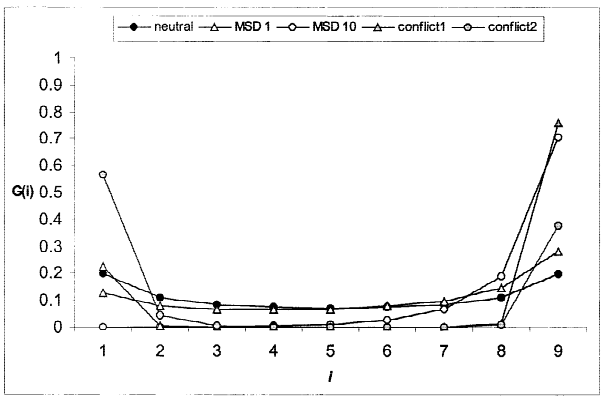

Fig. 2. The expected distributions of polymorphism under different models of evolution. The expected proportion of polymorphic sites at which the optimal codon is found at a frequency of *i* in a sample of ten sequences is given by G(*i*). The neutral model is equivalent to a MSD model (see Equations 2, 3) for which *S*, the measure of selection on the optimal codon at all sites, is zero. MSD 1 and MSD 10 represent models of evolution under which all sites are under the same selection coefficients of $S = 1$ and $S = 10$, respectively. The conflict models invoke two forms of selection, translation selection, S_t, and conflicting selection, S_c. For conflict model 1 we have $S_t = 200$ and $S_c = 500$, and for conflict model 2 we have $S_t = 200$, and $S_c = 250$ (G(*i*) values derived as shown in the Appendix).

tion to vary between genes (the move from gene specific selection coefficients to a single selection coefficient over all genes entails a reduction in *LogL* of 5.41), but the effect is not significant (the χ^2 approximation indicates that a reduction in *LogL* of 8.46 would be required for the gene-specific selection model to provide a significant improvement, at the 5% level, over the single selection model, given the difference of 9 degrees of freedom). A larger collection of genes may be required to demonstrate gene-specific selection coefficients.

We also estimated *S* from the average frequency of optimal codons at segregating sites, *K*, using Equation 4. The estimates derived from *K* are very similar to our maximum likelihood estimates based on the full polymorphism distribution (see Table 2, S_K values), suggesting that the data conform to the MSD model fairly well.

To assess the fit quality of the MSD model we used a *G*-test, which was performed on the combined spectrum (see Materials and Methods). The combined spectrum

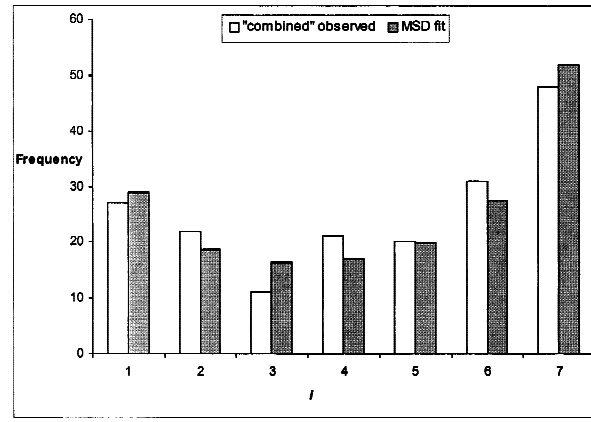

Fig. 3. The observed distribution of polymorphic synonymous sites (using the "combined" data—see Materials and Methods) compared with the distribution of polymorphic sites expected under the MSD model with the value of *S* fitted by ML. The observed distribution is the number of polymorphic sites at which the optimal codon is present at frequency *i* in the sample of eight sequences. The expected distribution is the expected proportion of polymorphic sites at which the optimal codon is found at a frequency of *i* in a sample of eight sequences (calculated according to Equations 2, 3), multiplied by 180, the total number of polymorphic synonymous sites.

yields a ML estimate of $S = 0.78$, which provides an adequate fit to the data ($G = 4.34$, $p = 0.50$, see Fig. 3).

Variation in Selection Between Amino Acids

Although we cannot reject the MSD model on the basis of its goodness of fit to the data, it is possible that other models may fit the data better. In particular, the pattern of synonymous codon bias in relation to gene expression levels suggests that selection is stronger on the codons of some amino acids than others (see Table 1, Eyre-Walker and Bulmer 1995; McVean and Vieira 1999). For some amino acids, like lysine, there is little or no change in the pattern of codon usage across expression levels, whereas for others, like glycine, there are dramatic changes (Eyre-Walker and Bulmer 1995). We used the data given by Sharp et al. (1992) to divide the amino acids into two groups: the codon usage of amino acids in the high optimal/sub-optimal group changes markedly with gene expression, while the codon usage of amino acids in the low optimal/sub-optimal group is relatively insensitive to gene expression (see Table 1).

The fixed site data show the expected trends: the high optimal/sub-optimal amino acids have 608 fixed optimal codons and 296 fixed sub-optimal codons, while the low optimal/sub-optimal amino acids have 932 fixed optimal codons and 638 fixed sub-optimal codons. Thus, the frequency of the optimal codon at fixed sites, *F*, is significantly higher for the high optimal/sub-optimal group of amino acids ($X^2 = 15.2$, $p = 0.0001$), which is consistent with stronger selection for codon usage in such amino acids.

The average frequency of optimal codons at segregat-

Table 3. The "combined" spectra of fixed and segregating sites

	Full "Combined"	Start of Genes	Rest of Genes	High Optimal/Sub-Optimal	Low Optimal/Sub-Optimal
0 (Sub-Optimal Fixed)	938	195	743	298	640
1	27	2	25	7	20
2	22	1	21	5	17
3	11	1	10	5	6
4	21	4	17	8	13
5	20	2	18	8	12
6	31	2	29	14	17
7	48	3	45	30	18
8 (Optimal Fixed)	1553	255	1298	610	943

ing sites, *K*, was significantly greater than 0.5 in the high optimal/sub-optimal group ($K = 0.65$, Wilcoxon signed rank, $p < 0.001$), but not significantly greater than 0.5 in the low optimal/sub-optimal group ($K = 0.52$, Wilcoxon signed rank, $p = 0.565$). Furthermore *K* was significantly different between the high and low optimal/sub-optimal groups (Mann-Whitney U test, $p < 0.0001$). The difference between high and low optimal/sub-optimal amino acids is also shown by the ML estimates of *S*, based on the combined spectra (see Table 3); for the high optimal/sub-optimal group $S = 1.96$ (95% confidence intervals 1.09 to 2.90), while for the low optimal/sub-optimal group $S = 0.00$ (95% confidence intervals −0.68 to 0.68).

Intragenic Variation in Selection

It has been suggested that conflicting selection pressures might be responsible for the reduction in codon bias at the start and ends of genes (Eyre-Walker 1996; Eyre-Walker and Bulmer 1993). In order to investigate this hypothesis, we separately analyzed the first 50 codons of the genes (the start) and all subsequent codons (the rest). In agreement with published results (Bulmer 1988; Chen and Inouye 1990; Eyre-Walker and Bulmer 1993) we find that the frequency of codons fixed for the optimal codon, *F*, is significantly lower at the start compared to the rest of the gene (see Table 3, $X^2 = 7.54$, $p = 0.006$). However, the frequency of the optimal codon at segregating sites, *K*, is not significantly higher at the start ($K = 0.56$ for the start versus $K = 0.58$ for the rest, Mann-Whitney U test, $p = 0.72$); and the ML estimates of *S* based on the combined spectra also fail to indicate significant intragenic differences ($S = 0.64$ for the start, 95% confidence intervals from −1.17 to 2.57, $S = 0.80$ for the rest, 95% confidence intervals from 0.26 to 1.35). These results are inconsistent with both the MSD and neutral models. Under the MSD model we expect similar results for fixed and segregating data, and *F* should positively covary with both *K* and *S*. Under the neutral model *F* can vary as the proportion of neutral sites varies, but *K* should be 0.5 and the estimated value of *S* should be zero.

However, the data are consistent with a conflicting selection pressures model in which a proportion of the sites are subject to strong conflicting selection pressures, and contribute little to polymorphism, while the rest of the sites are in a mutation-selection-drift balance. Under this conflict model the patterns of fixed and segregating sites can become uncoupled. Our data are thus qualitatively consistent with the hypothesis that conflicting selection pressures are responsible for the reduction in codon bias at the start of genes. We can test the conflict hypothesis further by looking at patterns of diversity: under this conflict model we predict a reduction in diversity at the start of genes, due to the effect of strong selection. In keeping with this prediction we find that the proportion of segregating sites is much lower for the start of genes than for the rest of genes (for the start of the genes there are 446 fixed and 17 segregating sites, for the rest of the genes there are 2028 fixed and 169 segregating sites, $X^2 = 9.51$, $p = 0.002$).

The Use of Fixed and Segregating Data

The evolutionary models discussed so far have invoked the infinite sites and unbiased mutation assumptions. While the infinite sites assumption appears to be justified for sites at which nucleotide diversity is low (Kimura 1971), and although the assumptions considerably simplify the analysis, there are some drawbacks. For instance, the infinite sites assumption restricts data analysis to segregating sites only. Thus, the infinite sites model does not allow a direct examination of the quantitative question of how many sites should be fixed for a sub-optimal codon. Furthermore, mutation may be biased with respect to optimal and sub-optimal codons (see Discussion).

The solution to these problems is to use the full Wright distribution to predict the frequencies of both fixed and segregating sites. The equations are similar to the case of infinite sites, except that *U* and *V* are no longer assumed to tend to zero (Equation 1), the integration of $F(x)$ is carried out between 0 and 1 (Equation 6) and we have used a Poisson probability model for the likelihood, summing from $i = 0$ to $i = n$ (Equation 7, see Hartl et al. 1994).

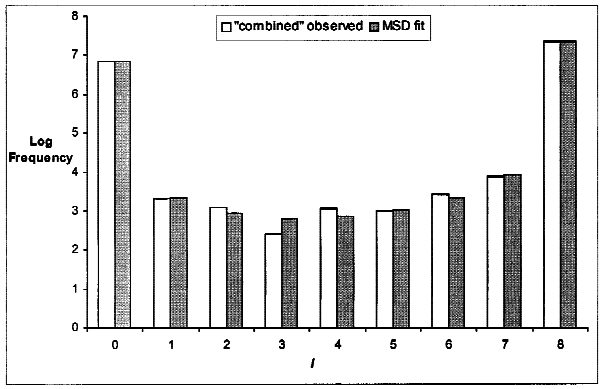

Fig. 4. The observed distribution of fixed and polymorphic synonymous sites (the "combined" data—see Materials and Methods) compared with the expected values obtained for the full Wright MSD model using the ML values of *S*, *U*, and *V*. The observed distribution is the number of fixed or polymorphic sites at which the optimal codon is present at frequency *i* in the sample of eight sequences. The expected distribution is given by G(*i*), the expected proportion of polymorphic sites at which the optimal codon is found at a frequency of *i* in a sample of eight sequences (calculated according to Equations 1, 9), multiplied by 2671, the total number of observed synonymous sites. The Y-axis is plotted on a log scale to aid visualization.

$$F(x) = C\,E^{Sx}x^{(V-1)}(1-x)^{(U-1)} \tag{1}$$

$$G(i) = C\int_{x=0}^{1} \frac{n!}{i!(n-i)!}\,x^{i}(1-x)^{n-i}\,F(x)dx \tag{6}$$

$$LogL = \sum_{i=0}^{n} Log\left[\frac{E^{-N(i)}\,G(i)^{N(i)}}{N(i)!}\right] \tag{7}$$

We can use the full Wright distribution to see whether the MSD model provides a reasonable fit to both the fixed and the segregating data. Although the infinite sites assumption is no longer required, we shall again consider all genes except *gnd*, in order to compare the two approaches. The use of the full Wright distribution requires that we estimate the three parameters *S*, *U*, and *V*. The combined spectrum yields estimates of $S = 0.82$, $U = 0.032$, and $V = 0.024$. This result is supportive of the infinite sites and unbiased mutation assumptions, since the estimate of *S* is similar to the value obtained using the infinite sites assumption on the combined polymorphism data ($S = 0.78$). The MSD model provides a reasonable fit to the data (see Figure 4), as confirmed by a *G* test ($G = 4.34$, $p = 0.50$).

We can examine the evidence for gene specific evolutionary parameters by specifying a number of MSD models with different numbers of parameters, and then comparing their fits to the data using the χ^2 approximation to $2\Delta LogL$. The models are compared on the basis of the full data of all seven genes for which there is sufficient data to estimate three parameters (*crr*, *mdhA*, *pabB*, *phoA*, *putP*, *sppA*, and *zwf*). MSD model 1 has all genes with the same *S*, *U*, and *V* parameter values (3 parameters). MSD model 2 has all genes sharing the same *U* and *V* values but allows gene specific values of *S* (9 parameters). MSD model 3 has all genes sharing the same *S* value but allows gene specific values of *U* and *V* (15 parameters). MSD model 4 allows gene specific values for *S*, *U*, and *V* (21 parameters).

As expected the more complex the model the better the fit to the data, as shown by the maximum *LogL* values increasing from MSD model 1 to MSD model 4 (MSD model 1 $LogL = -202.722$, MSD model 2 $LogL = -177.155$, MSD model 3 $LogL = -159.033$, MSD model 4 $LogL = -154.184$). We can compare those models which are nested within one another: this means we can perform all pairwise model comparisons except MSD model 2 versus MSD model 3. When compared with MSD model 1, all three alternative models provide a significantly better fit to the data (model 2 $p = 2 \times 10^{-9}$, model 3 $p = 2 \times 10^{-13}$, model 4 $p = 8 \times 10^{-13}$). When we compare MSD model 2 with MSD model 4, we find that model 4 provides a significantly better fit to the data ($p = 7 \times 10^{-6}$). But when we compare MSD model 3 with MSD model 4, we find that the improved fit of model 4 does not justify the extra parameters required ($p = 0.138$). This means that MSD model 3 is the best MSD model of those considered, taking into account the number of parameters required. Thus, we have significant evidence of variation in mutation, but the evidence of variation in selection is non-significant, as we also found using polymorphic data only.

Since we need to combine the data from many genes for accurate parameter estimation, these results suggest that the use of the polymorphic data under the infinite sites assumption may be preferable to the use of the fixed and polymorphic data using the full Wright distribution, since then variation in mutation rates can be ignored. Although the full Wright distribution avoids the infinite sites assumption and allows the use of additional fixed site data, its use does require the estimation of many additional parameters.

Discussion

Although some synonymous codons appear to be favored by natural selection over others, they are not used at every site in every gene. We have tested three hypotheses for why translationally sub-optimal codons exist: (1) there is a balance between mutation, selection and genetic drift (the MSD model), (2) selection is absent at certain codons (the neutral model), and (3) sub-optimal codons are favored at some sites by alternative, conflicting selection pressures (the conflict model). We have attempted to differentiate between these hypotheses by considering frequency distributions of translationally optimal codons in *E. coli*.

Codon usage in E. coli *supports the MSD hypothesis*

WE find that on average optimal codons segregate at higher frequencies than sub-optimal codons. The MSD model with a single selection coefficient fits the data well, and this result suggests that sub-optimal codons persist primarily because selection is weak relative to drift. The polymorphism data are inconsistent with the neutral and conflict models as sole explanations for the persistence of sub-optimal codons: under the neutral model we expect sub-optimal and optimal codons to segregate at similar frequencies, and under most conflict models we expect all polymorphisms to segregate at very low frequencies with a strong bias towards the optimal or sub-optimal codon being rare, depending on the specific model.

Evidence for the Neutral and Conflict Hypotheses Applying at Some Sites

However, we do find some evidence for the neutral and conflict hypotheses in sub-sets of the data. It has been noted previously that some amino acids such as lysine show little change in codon usage across expression levels (Eyre-Walker and Bulmer 1995; McVean and Vieira 1999), suggesting selection on codon usage is absent or weak for those codons. Our results confirm these observations, since amino acids which show little variation in codon usage across expression levels show frequency distributions consistent with neutrality (for the low optimal/sub-optimal group of amino acids we estimate S to be 0.00 with 95% confidence intervals of −0.68 and 0.68; however note that possible biases discussed below will probably cause S to be underestimated).

It has also been noted that codon bias is lower at the start of enterobacterial genes (Bulmer 1987; Chen and Inouye 1990; Eyre-Walker and Bulmer 1993; Hartl et al. 1994). This could be due to weaker translational selection at the start of genes or conflicting selection pressures. If it was the former we would expect the average frequency of the optimal codon at segregating sites to be lower at the start of the gene, and the resulting strength of selection estimated from the polymorphism data to be lower. This we do not find since the average frequency of the optimal codon at segregating sites and the estimated selection strength are very similar at the start of the genes and in the rest of the genes. However, the overall level of diversity is significantly lower at the start of the genes and this suggests that strong conflicting pressures are acting at some sites, while others are evolving in a mutation-selection-drift balance. It has been suggested that these conflicting selection pressures might be associated with selection to regulate gene expression by the use of sub-optimal codons at the start of the gene (Chen and Inouye 1990), selection to avoid secondary structure, or selection upon particular ribosome binding motifs within genes (Eyre-Walker and Bulmer 1993). The results presented here do not help us differentiate between these possibilities.

A Consideration of the Assumptions of the Population Genetics Model

We have used a simple population genetics model in order to discriminate between alternative hypotheses for the existence of sub-optimal codons. The simple model requires a number of assumptions which may not apply to the *E. coli* data we have analyzed. Here we consider five important assumptions: equilibrium, unbiased mutation, random sampling, constant evolutionary parameters, and independence of sites.

Equilibrium

We assume that mutation, selection, and population size have been constant for long enough so that all evolutionary processes are at equilibrium. We do not test the non-equilibrium hypothesis in this paper, although we note that synonymous codon bias appears to be relatively stable in enteric bacteria suggesting that a state of equilibrium exists (Maynard Smith and Smith 1996), that the levels of codon bias in *E. coli* and *S. typhimurium* are similar (Sharp 1991), and the rate of synonymous substitution is considerably greater than the rate of non-synonymous substitution, suggesting that any effects of amino acid substitution will be short lived (Sharp 1991).

Unbiased Mutation

In the majority of the paper we use the infinite sites model, which means that mutation rates are very low, and we also assume unbiased mutation, that there is no tendency for the mutation rate from an optimal codon to a sub-optimal codon to be greater or less than the mutation rate in the opposite direction. While this assumption may appear reasonable, a possible mutational bias could result from compositional correlations. Most optimal codons end in G or C (16 out of 25, see Table 1), while most sub-optimal codons end in A or T (21 out of 34, see Table 1). So if there are mutational biases with regard to composition, there are probably mutational biases with regard to codon usage.

The assumption of unbiased mutation can be justified on the basis that when mutation rates are low, mutation is effectively unbiased, even if U/V is not precisely unity. When the assumption of unbiased mutation rates and infinite sites is relaxed with the use of the full Wright distribution (Equation 1) rather than the simplified form (Equation 2), we find no qualitative effect on our results. In particular, we can be sure that our support of the MSD hypothesis over the neutral hypothesis is not an artefact of our assumption of unbiased mutations rates.

Furthermore, the effect of mutation bias on the distribution of optimal codon frequencies is very weak when mutation rates are low. If we compare the expected fre-

quency distributions at fixed and segregating sites for contrasting strong mutational biases under weak selection, we find the differences in frequency between the two mutational extremes to be low (between 1% and 7%, depending on the frequency class, using Equations 1 and 9, with $n = 10$ and $S = 1$ in both cases, and either $U = 0.02$ and $V = 0.04$ or $U = 0.02$ and $V = 0.04$). So although we may expect $K > 0.5$ under neutrality if mutation rates are biased, the deviation is expected to be very small (using Equations 1 and 9, with $n = 10$, $S = 0$, $U = 0.02$ and $V = 0.04$, we obtain $K = 0.506$).

Random sampling

We assume that samples are obtained at random from the population. This assumption is unlikely to met, with sampling strategies likely to lead to overdispersion of samples (see primary literature cited in Materials and Methods). The effect of such overdispersed sampling on the frequency distribution of polymorphism is unclear, since it is dependent on the shape of the phylogeny underlying the sampling, but the effect is unlikely to be extreme since the MSD model provides a reasonably good fit to the polymorphism data (see Figure 3).

Constant evolutionary parameters

We assume that the selection and mutation parameters, *S*, *U*, and *V*, do not vary between sites. If the evolutionary parameters do vary between sites, as often appears to be the case (Yang 1996), then estimates based on the assumption of no variation may be biased.

Independence of sites

Sites can evolve independently if there is no epistasis and free recombination. However, the latter is unlikely to apply in *E. coli*, and so we must consider the possible effects of selection interference. In general, selection at one site will reduce the efficacy of selection at linked sites (Hill and Robertson 1966). All forms of selection can cause selection interference: strong positive selection causes hitchhiking (Gillespie 2000; Maynard Smith and Haigh 1974), strong negative selection causes background selection (Charlesworth et al. 1993), and weak selection causes weak selection interference (Comeron et al. 1999; Li 1987; McVean and Charlesworth 2000).

The effect of selection interference will be to cause selection coefficients based on the frequency distribution of optimal codons to be underestimated. Weak selection interference will change the frequency distribution of optimal codons in very nearly the same way as a reduction in the selection coefficient. Background selection will make little difference to the frequency distribution other than through a reduction in the effective population size (strongly deleterious mutations only reach low frequencies before removal by selection). Hitchhiking will increase the proportion of singleton polymorphisms in addition to reducing the effectiveness of selection, but as long as we assume that hitchhiking events are independent of synonymous codon usage, the additional effect on the inferred level of selection should be small.

The Paradox of Similar Codon Usage Across Species and Genes

Our results suggest that a large proportion of sub-optimal codons are maintained because they are held in a balance between mutation, selection, and genetic drift. However, this presents us with a puzzle. The MSD model is parameter sensitive: if $N_{eS} << 1$ then selection has little effect and synonymous codon use is determined by mutation bias, and if $N_{eS} >> 1$ then selection is so strong that we expect optimal codons at every site. Yet, synonymous codon bias is similar across a broad range of bacteria, which might be expected to have different population sizes, and also across different genes, which might be expected to have different synonymous selection coefficients.

Let us first consider the paradox of similar codon usage across species. Interference at linked selected sites, either due to weak selection interference or hitchhiking, appears a promising solution to this problem. The MSD model is less parameter sensitive if many weakly selected sites are linked together because the mutations tend to interfere with each other in such a way as to reduce the effects of variation in census population size; as the census population size increases, interference increases also, leading to a less than proportional increase in the effective population size (McVean and Charlesworth 2000). With low rates of recombination, as expected in bacteria, the effect can be dramatic with little change in the observed level of codon bias over changes in the census population size of several orders of magnitude (McVean and Charlesworth 2000).

The apparent constancy in the degree of codon bias across species could also be due to hitchhiking. If there is no recombination then the probability that a weakly selected segregating mutation will be swept to fixation depends on whether it is on the chromosome on which the new mutation occurs. If advantageous substitutions are frequent enough then the major stochastic force acting upon mutations of small selective effect is not genetic drift, but *genetic draft* (Gillespie 2000), the probability that they will be linked to the advantageous mutations. If adaptive evolution is mutation limited, then as the population size increases, the number of advantageous substitutions increases and genetic draft becomes a more potent force, reducing the efficacy of selection at the weakly selected sites. The details of this model have not been worked out for weakly selected mutations, but it is anticipated that the continual fixation of advantageous mutations will tend to uncouple the rate of evolution for weakly selected mutations from the census population size, just as it uncouples the level of neutral genetic variation (Gillespie 2000).

The paradox of similar codon usage across genes is raised by simple biochemical models which suggest that the strength of translational selection upon codon usage bias should be proportional to gene expression level, whether the selection is acting upon the rate of elongation, the cost of proofreading, or translational accuracy (Bulmer 1991). The levels of gene expression, as measured by the number of protein molecules in an *E. coli* cell, vary by several orders of magnitude, from proteins such as the *lac* repressor which is present at a concentration of ten molecules per cell, to the product of the *ompA* gene, present at a concentration of 36,000 molecules per cell (see Eyre-Walker 1996). *lacI* has minimal codon bias, but genes such as *leuS* have significant codon bias and are found at concentrations of 500 molecules per cell. Why, then, are there sub-optimal codons in the *ompA* gene where the strength of selection is expected to be 72 times stronger than in *leuS*, where the strength of selection is strong enough to cause codon bias? There are a number of possible reasons. FIrst, the biochemical models developed by Bulmer (1991) may be incorrect and the strength of selection on synonymous codon use may not be proportional to gene expression level; in particular Bulmer assumed that fitness was proportional to growth rate, and this may not be the case for *E. coli* growing in its natural environment where its doubling rate is substantially slower than in the laboratory (Savageau 1983). Second, the differences in expression level, which are measured in exponentially growing cells may not reflect the differences in expression level in naturally growing *E. coli.* And finally, weak selection interference may explain the data if there is sufficient recombination between genes to uncouple the interference experienced by one gene from that of others.

Many years ago it was suggested that synonymous sites may be a paradigm of neutral evolution (King and Jukes 1969). In a strict sense that has turned out not to be the case, since selection undoubtedly operates on synonymous codon use in many organisms. However the present results suggest that stochastic processes have had a major impact on the evolution of synonymous codon use in *E. coli.*

Acknowledgments. Thanks to the BBSRC (NGCS) and the Royal Society (AE-W) for support.

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*—natural selection and translational accuracy. Genetics 136: 927–935

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. Genetics 139:1067–1076

Akashi H (1996) Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics 144: 1297–1307

Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8:688–693

Akashi H, Schaeffer SW (1997) Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics 146:295–307

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

Boyd EF, Nelson K, Wang F, Whittam TS, Selander RK (1994) Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica.* Proc Natl Acad Sci USA 91:1280–1284

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

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

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

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303

Chen GT, Inouye M (1990) Suppression of the negative effect of minor arginine on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. Nucl Acids Res 18:1465–1473

Comeron JM, Kreitman M, Agaude M (1999) Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics 151:239–249

DuBose RF, Dykhuizen DE, Hartl DL (1988) Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli.* Proc Natl Acad Sci USA 85:7036–7040

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

Ewens WJ (1979) Mathematical population genetics. Springer-Verlag, Berlin

Eyre-Walker A (1996) Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol 13:864–872

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

Eyre-Walker A, Bulmer M (1995) Synonymous substitution rates of enterobacteria. Genetics 140:1407–1412

Gillespie JH (2000) Genetic drift in an infinite population: the pseudohitchhiking model. Genetics 155:909–919

Gouy G, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Guttman DS, Dykhuizen DE (1994a) Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266:1380–1383

Guttman DS, Dykhuizen DE (1994b) Detecting selective sweeps in naturally occurring *Escherichia coli.* Genetics 138:993–1003

Hall BG, Sharp PM (1992) Molecular population genetics of *Escherichia coli*: DNA sequence diversity at the *celC*, *crr* and *gutB* loci of natural isolates. Mol Biol Evol 9:654–665

Hartl DL, Moriyama EN, Sawyer SA (1994) Selection intensity for codon bias. Genetics 138:227–234

Hill WG, Robertson A (1966) The effect of linkage on limits to natural selection. Genet Res 8:269–294

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

Kimura M (1971) Theoretical foundation of population genetics at the molecular level. Theor Popul Biol 2:174–208

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Li WH (1987) Models of nearly neutral mutations with particular im-

plications for nonrandom usage of synonymous codons. J Mol Evol 24:337–345
Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. Genet Res 23:23–35
Maynard Smith J, Smith NH (1996) Site-specific codon bias in bacteria. Genetics 142:1037–1043
McVean GAT, Charlesworth B (2000) The effects of Hill-Robertson interference between weakly selected sites on patterns of molecular evolution and variation. Genetics 155:929–944
McVean GAT, Vieira J (1999) The evolution of codon preferences in Drosophila: a maximum-likelihood approach to parameter estimation and hypothesis testing. J Mol Evol 49:63–75
Nelson K, Selander RK (1992) Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. J Bacteriol 174:6886–6895
Nelson K, Whittam TS, Selander RK (1991) Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. Proc Natl Acad Sci USA 88:6667–6671
Savageau MA (1983) *Escherichia coli* habitats, cell types and molecular mechanisms of gene control. Am Nat 122:732–744
Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132:1161–1176
Sawyer SD, Dykhuizen DE, Hartl DL (1987) Confidence interval for the number of selectively neutral amino-acid polymorphisms. Proc Natl Acad Sci USA 84:6225–6228
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, Burgess CJ, Lloyd AT, Mitchell KJ (1992) Selective use of termination codons and variation in codon choice. In: Hatfield DL, Lee BJ, Pirtle RM (eds) Transfer RNA in protein synthesis. CRC Press, Boca Raton, pp 397–425
Vogel RF, Entian KD, Mecke D (1987) Cloning and sequence of the *mdh* structural gene of *Escherichia coli* coding for malate dehydrogenase. Arch Microbiol 149:36–42
Wright S (1949) Adaptation and selection. In: Jepson G, Simpson G, Mayr E (eds). Genetics, paleontology and evolution. Princeton University Press, Princeton, pp 365–391
Yang Z (1996) The among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372

Appendix

We can generalize Equation 3 to account for m different classes of sites which contribute to a proportion α_j of segregating polymorphisms and at which the optimal codon is favored by the selective coefficient S_j. Then $G(i)$ can be written as

$$G(i) = C \sum_{j=1}^{m} \alpha_j C_j \int_{x=1/Ne}^{1-1/Ne} \frac{n!}{i!(n-1)!} x^i (1-x)^{n-i} F(x,S_j) dx \tag{A1}$$

C_j is a constant which ensures that the integral of $F(x,S_j)$ is one. Thus, the conflict model parameterises as $m = 3$, $\alpha_1 = ½\ (Q_1/Q)$, $S_1 = S_t$; $\alpha_2 = ¼\ (Q_2/Q)$, $S_2 = S_t + S_c$; $\alpha_3 = ¼\ (Q_3/Q)$, $S_3 = S_t - S_c$; $Q = Q_1 + Q_2 + Q_3$. The Q parameters are scaling factors which convert from proportions of all sites, both segregating and fixed, to proportions of segregating sites only. We can derive the Q parameters using the full Wright distribution (Equation 1 in the main text and Equation A2 below) from which we can derive the distribution of both fixed and segregating sites according to Equation A3. U and V are not considered to be zero, but rather reflect very low and unbiased mutation rates which cause very little deviation from the infinite sites assumption ($U = V = 10^{-6}$).

$$F(x,S) = C_F E^{Sx} x^{(V-1)} (1-x)^{(U-1)} \tag{A2}$$

$$G(i,S) = C_G \int_{x=0}^{1} \frac{n!}{i!(n-i)!} x^i (1-x)^{n-i} F(x,S) dx \tag{A3}$$

In the case of fixed and segregating sites $G(i)$ is defined from $i = 0$ to $i = n$. C_F is the normalizing constant which ensures that the integral of $F(x,S)$ between $x = 0$ and $x = 1$ is unity, and C_G is the normalizing constant which ensures that the sum of $G(i,S)$ between $i = 0$ and $i = n$ is unity. From Equation A3 we can determine the proportion of segregating sites, $Q(S)$.

$$Q(S) = \sum_{i=1}^{n-i} G(i,S) \tag{A4}$$

In the conflict model we have three classes of sites, hence $m = 3$, for which we know the selective coefficient, S_j, and the proportions at all sites, P_j. We can now determine α_j, the proportions at segregating sites.

$$\alpha_j = \frac{Q(S_j)P_j}{\sum_{j=1}^{m} Q(S_j)P_j} \tag{A5}$$

In conflict model 1 we have $S_t = 200$ and $S_c = 500$. So we know the proportions of all sites ($P_1 = ½$, $P_2 = ¼$, and $P_3 = ¼$), and the selection coefficients ($S_1 = 200$, $S_2 = 700$, and $S_3 = -300$). With $n = 10$, we obtain $\alpha_1 = 0.677$, $\alpha_2 = 0.097$, and $\alpha_3 = 0.226$, which yield the distribution of polymorphism shown in Fig. 2.