J Mol Evol (1986) 24:28–38

Journal of
Molecular Evolution

An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms*

Paul M. Sharp[1,2] and Wen-Hsiung Li[1]

[1] Center for Demographic and Population Genetics, University of Texas, PO Box 20334, Houston, Texas 77225, USA
[2] Department of Genetics, Trinity College, Dublin 2, Ireland

Summary. Observed patterns of synonymous codon usage are explained in terms of the joint effects of mutation, selection, and random drift. Examination of the codon usage in 165 *Escherichia coli* genes reveals a consistent trend of increasing bias with increasing gene expression level. Selection on codon usage appears to be unidirectional, so that the pattern seen in lowly expressed genes is best explained in terms of an absence of strong selection. A measure of directional synonymous-codon usage bias, the Codon Adaptation Index, has been developed. In enterobacteria, rates of synonymous substitution are seen to vary greatly among genes, and genes with a high codon bias evolve more slowly. A theoretical study shows that the patterns of extreme codon bias observed for some *E. coli* (and yeast) genes can be generated by rather small selective differences. The relative plausibilities of various theoretical models for explaining nonrandom codon usage are discussed.

Key words: Codon usage — Synonymous substitution rate — Codon Adaptation Index — Enterobacterial genes — G + C content — Theoretical models

Introduction

Since synonymous mutations cause no change in gene products, they have commonly been thought to be subject to few selective constraints and have been considered by some evolutionists (Kimura 1968; King and Jukes 1969) to be good candidates for selectively neutral mutations. For this reason, unequal usage of the alternative codons for an amino acid was not anticipated. However, with the determination of a substantial number of DNA sequences, it became apparent that nonrandom usage of synonymous codons is a general phenomenon (Grantham et al. 1980, 1981). There now exists a large body of codon usage data, from which two general observations have been made:

1. Genes within a species usually have similar patterns of codon preference, but genes from different taxonomic groups have different patterns (Grantham et al. 1980). For example, *Escherichia coli* and *Salmonella typhimurium* (two closely related enteric bacteria) show very similar preferences (Ikemura 1985), but the unrelated bacterium *Bacillus subtilis* shows quite different preferences (Ogasawara 1985).

2. Despite observation 1, considerable heterogeneity exists in codon usage patterns within species. In unicellular organisms, highly expressed genes exhibit a greater degree of bias in favor of a particular subset of codons than do lowly expressed genes (Bennetzen and Hall 1982; Gouy and Gautier 1982). In mammalian genomes, which seem to be mosaics of regions of rather different G + C content (Bernardi et al. 1985), codon usage in any particular gene seems to be related to the degree of local GC richness (Bernardi and Bernardi 1985; Ikemura 1985).

From a theoretical viewpoint an observed pattern of codon usage reflects the joint action of mutation and natural selection, but what is the relative importance of these two forces? The importance of

Offprint requests to: W.-H. Li
* Presented at the FEBS Symposium on Genome Organization and Evolution, held in Crete, Greece, September 1–5, 1986

natural selection in forging the highly biased pattern of codon usage in highly expressed genes in *E. coli* and yeast was strongly suggested by two lines of work involving tRNAs. First, the relative abundances of different tRNA species vary, and those sets of codons translated by the more abundant species are used more frequently (Ikemura 1981a,b, 1982). Second, analysis of tRNA anticodon sequences shows that within a set of codons recognized by the same tRNA, those that might be expected to form the optimal codon–anticodon interaction are more frequently used (Bennetzen and Hall 1982; Grosjean and Fiers 1982; Ikemura 1985). Thus it appears that selection at the level of translation has heavily favored certain "optimal" codons in genes expressed at high levels. Interestingly, the tRNA abundance profiles and the anticodon sequences differ between *E. coli* and yeast, and so do the optimal codons; this may largely explain observation 1.

The relative importance of mutation pressure in determining codon usage has not been established. In prokaryotes, where there is very little superfluous DNA, there is a correlation between G + C content at synonymous sites in codons and in the genome as a whole (Bibb et al. 1985), but it is not clear how the correlation has arisen. There is evidence that differences in genomic G + C content could arise from differences in the spontaneous rates of the different possible base substitutions. For example, the comparative A + T richness of the genome of *Herpesvirus saimiri* may be related to the presence of a gene for thymidylate synthase, which is lacking in several herpesviruses that are G + C rich (Honess et al. 1986). Since genomic G + C content varies considerably among organisms, this correlation of G + C content and codon usage would also contribute to the taxon-specific pattern of codon usage. In mammals there is a strong relation between G + C content at synonymous sites and in neighboring introns (Ikemura 1985), and less evidence of selection.

Our understanding of the phenomenon of nonrandom codon usage has greatly increased in recent years but remains rather incomplete. In this paper we examine this phenomenon from an evolutionary perspective. First, we show that in *E. coli* there is a consistent trend in synonymous-codon usage bias, from a very high bias in highly expressed genes to a low bias in lowly expressed genes. This strongly suggests that the pattern of codon usage in highly expressed genes is determined largely by selection, whereas in lowly expressed genes mutation and random drift are also influential. We refute the suggestion, made on numerous occasions (e.g., Grosjean and Fiers 1982; Konigsberg and Godson 1983; Hinds and Blake 1985), that the relatively high incidence in lowly expressed genes, particularly regulatory ones, of certain codons recognized by minor tRNA species represents an evolutionary strategy used to lower the level of gene expresion. Rather, we suggest that the higher incidence is due simply to a relaxation of selection in lowly expressed genes. Second, we show that a negative correlation exists between the degree of codon usage bias and the rate of synonymous substitution in a gene. From this we conclude that variation in the degree of codon usage bias and variation in the rate of synonymous substitution among genes are two aspects of the same phenomenon—they both reflect variation among genes in the selective constraints on synonymous changes. Third, we address the question of the magnitude of selective difference required to produce a strong codon usage bias, and discuss the relative plausibilities of various theoretical models for explaining nonrandom codon usage.

Evidence for a Unidirectional Trend in Codon Preference

We suggest that in unicellular organisms there is a single trend in synonymous codon usage, from a high bias in highly expressed genes (where selection on codon usage is strong) to a low bias in lowly expressed genes (where selection is weak).

To support this assertion we compiled codon usage data for 165 *E. coli* chromosomal genes (Sharp and Li 1986). Ideally, we would then have categorized these genes by expression level, but such data are not readily available for all the genes. Furthermore, it is not clear whether the constitutive expression level or a transient maximum expression level is more important (Gouy and Gautier 1982). However, we did extract several subsets of genes. They were categorized as "very highly expressed" (27 genes, mainly encoding ribosomal proteins, elongation factors, and outer membrane proteins), "highly expressed" (15 genes, including those encoding RNA polymerase subunits and aminoacyl tRNA synthetases), and "regulatory" (8 genes, encoding regulatory or repressor proteins and expressed at very low levels).

This left a heterogeneous group of "others" (115 genes of very mixed expression level). This group was divided into groups of "moderate" and "low" codon bias on the basis of usage of seven particular pairs of codons, namely those pairs of synonymous codons that are A/U or G/C rich at codon positions one and two, and have a pyrimidine (Y = U or C) at position three [the pair CCY was excluded for reasons given in Sharp and Li (1986)]. A preference for U in G/C-rich codons and C in A/U-rich codons has been observed in highly expressed genes in both *E. coli* (Gouy and Gautier 1982) and yeast (Sharp et al. 1986). Although this phenomenon is not com-

Table 1. Relative synonymous-codon usage (RSCU) values in 165 *E. coli* genes

		Gene group							Gene group				
		VH	H	M	L	R			VH	H	M	L	R
Phe	UUU	0.46	0.60	0.72	1.11	1.30	Ser	UCU	2.57	1.75	1.32	0.83	0.80
	UUC	1.54	1.40	1.28	0.89	0.70		UCC	1.91	1.75	1.35	0.83	0.85
Leu	UUA	0.11	0.17	0.39	0.74	0.88		UCA	0.20	0.26	0.55	0.59	0.89
	UUG	0.11	0.36	0.55	0.79	0.81		UCG[a]	0.04	0.48	0.84	0.95	0.93
Leu	CUU	0.22	0.33	0.49	0.54	0.77	Pro	CCU	0.23	0.47	0.45	0.55	0.54
	CUC	0.20	0.45	0.57	0.64	0.49		CCC[a]	0.04	0.07	0.15	0.52	0.81
	CUA[a]	0.04	0.07	011	0.18	0.11		CCA	0.44	0.50	0.72	0.75	0.71
	CUG	5.33	4.62	3.89	3.12	2.94		CCG	3.29	2.97	2.68	2.19	1.95
Ile	AUU	0.47	0.96	1.14	1.64	1.56	Thr	ACU	1.80	0.97	0.77	0.62	0.48
	AUC	2.53	2.03	1.78	1.24	1.13		ACC	1.87	2.37	2.06	1.78	1.93
	AUA[a]	0.01	0.01	0.08	0.12	0.31		ACA	0.14	0.13	0.33	0.48	0.41
Met	AUG	1.00	1.00	1.00	1.00	1.00		ACG	0.18	0.53	0.84	1.13	1.17
Val	GUU	2.24	1.51	1.23	0.98	1.08	Ala	GCU	1.88	0.93	0.79	0.53	0.50
	GUC	0.15	0.53	0.69	0.89	1.18		GCC	0.23	0.68	0.88	1.24	1.15
	GUA	1.11	0.88	0.65	0.60	0.39		GCA	1.10	0.92	0.86	0.74	0.70
	GUG	0.50	1.09	1.43	1.53	1.35		GCG	0.80	1.47	1.48	1.49	1.65
Tyr	UAU	0.39	0.67	0.91	1.18	1.13	Cys	UGU	0.67	0.76	0.87	0.79	1.14
	UAC	1.61	1.33	1.09	0.82	0.87		UGC	1.33	1.24	1.13	1.21	0.86
ter	UAA	—	—	—	—	—	ter	UGA	—	—	—	—	—
ter	UAG	—	—	—	—	—	Trp	UGG	1.00	1.00	1.00	1.00	1.00
His	CAU	0.45	0.57	0.76	1.14	1.12	Arg	CGU	4.39	3.86	3.33	2.17	1.90
	CAC	1.55	1.43	1.24	0.86	0.88		CGC	1.56	2.00	2.16	2.76	2.70
Gln	CAA	0.22	0.35	0.54	0.66	0.80		CGA[a]	0.02	0.06	0.18	0.29	0.54
	CAG	1.78	1.65	1.46	1.34	1.20		CGG[a]	0.02	0.03	0.21	0.57	0.65
Asn	AAU	0.10	0.35	0.54	0.91	1.13	Ser	AGU	0.22	0.24	0.43	0.87	0.76
	AAC	1.90	1.65	1.46	1.09	0.87		AGC	1.05	1.52	1.51	1.93	1.78
Lys	AAA	1.60	1.45	1.53	1.51	1.45	Arg	AGA[a]	0.02	0.02	0.09	0.13	0.11
	AAG	0.40	0.55	0.47	0.49	0.55		AGG[a]	0.00	0.03	0.04	0.09	0.11
Asp	GAU	0.61	0.94	1.09	1.28	1.26	Gly	GGU	2.28	2.23	1.80	1.34	1.31
	GAC	1.39	1.06	0.91	0.72	0.74		GGC	1.65	1.50	1.67	1.74	1.55
Glu	GAA	1.59	1.45	1.44	1.37	1.30		GGA[a]	0.02	0.08	0.20	0.33	0.39
	GAG	0.41	0.55	0.56	0.63	0.70		GGG[a]	0.04	0.19	0.33	0.59	0.75

Data from Sharp and Li (1986). Groups of genes (see text for description): VH, 27 very highly expressed genes (total 6240 codons); H, 15 highly expressed genes (9223 codons); M, 57 genes with moderate codon bias (47,622 codons); L, 58 genes with low codon bias (22,612 codons); R, 8 regulatory or repressor genes (2462 codons)

[a] "Rare" codons, defined by RSCU < 0.05 in the VH group

pletely understood, for UUY, UAY, and AAY the preference may result from selection against wobble-type pairing, as the cognate tRNAs all have G at the first position of the anticodon (Fitch 1976). A statistic to measure this preference, P2′, was modified (Sharp and Li 1986) from that used by Gouy and Gautier (1982). The mean P2′ value for the very highly expressed genes is 0.76 (range, 0.62–0.92), and that for the highly expressed genes is 0.65 (range, 0.52–0.82). The P2′ values for the 115 "other" genes range between 0.30 and 0.76, and we selected the upper half of this distribution (57 genes with P2′ > 0.49) as the group with moderate codon bias. The group of genes with P2′ < 0.49 were designated the low-codon-bias group.

The compiled codon usage data for the five groups of genes are presented in Table 1. To enable comparison between data sets of different sizes, the codon usage numbers are converted into relative synonymous-codon usage (RSCU) values. The RSCU value for a codon is simply the observed frequency of that codon divided by the frequency expected under the assumption of equal usage of the synonymous codons for an amino acid (Sharp et al. 1986). Thus,

$$\mathrm{RSCU}_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}} \qquad (1)$$

where X_{ij} is the number of occurrences of the j-th codon for the i-th amino acid, and n_i is the number (from 1 to 6) of alternative codons for the i-th amino acid.

In Table 1 it can be seen that the two groups of highly expressed genes differ to some extent in codon usage. Generally the same codons are preferred or disfavored in the two groups, but the bias is stronger in the very highly expressed group. The trend con-

tinues through the moderate- and low-codon-bias groups, separated on the basis of bias in a particular subset of codons comprising less than a quarter of the whole code. An important implication of this consistent trend is that for all amino acids the codon usage bias seems to be influenced by the same common factors. Of these the most important is probably the level of expression, as has been suggested by many authors (e.g., Grantham et al. 1981; Ikemura 1981b; Grosjean and Fiers 1982).

A clear example of the directional trend in codon preference is seen in the usage of Asn codons, where a very strong bias in favor of AAC in very highly expressed genes declines progressively to a very weak bias in the low-codon-bias group. Of course, the two Asn codons contribute to the P2′ statistic, and so this situation arises partly because of the criteria used to divide the non-highly expressed genes. However, a similar, though less marked, trend in the usage of Gln codons is independent of the P2′ statistic. For the quartet of Pro codons there is a consistent decrease (again independent of P2′) in the bias in favor of CCG, while the three other codons not only increase in relative frequency, but also tend toward uniformity of frequency. A similar pattern emerges for the six Leu codons, although even in the low-codon-bias group CUG remains much favored and CUA rather rare.

The trend in codon bias as one moves across the groups of genes away from the very highly expressed genes is clearly toward more uniform use of alternative synonymous codons. However, this trend is not expected to be toward precisely equal use of all synonymous codons, since in very lowly expressed genes the pattern of usage would be largely determined by mutation pressure. Data pertaining to the pattern of spontaneous mutation are scarce for *E. coli* (see, e.g., Schaaper et al. 1986). However, extensive data have been gathered from mammalian pseudogenes, where mutation rates do not drive the four nucleotides toward equal frequencies (Li et al. 1984) and where neighboring-base effects on patterns of mutation have also been detected (Bulmer 1986). In Table 1, for several amino acids (His and Asp are clear examples) the bias in favor of one codon declines to the extent of becoming reversed; i.e., the rarer codon in very highly expressed genes becomes the more common codon in the low-codon-bias group. Interestingly, in several cases where the codon preference switches in the low-codon-bias group the pattern seems to reflect dinucleotide frequencies in the *E. coli* genome as a whole. Thus, AT appears to be more frequent than AC in *E. coli* DNA (Nussinov 1984), and in the low-codon-bias group of genes His and Asp are more frequently encoded by CAU (in preference to CAC) and GAU (in preference to GAC), respectively.

The above observations can be simply explained by assuming that selection (against nonoptimal codons) becomes weaker as gene expression decreases. Of course, from observations of codon frequencies we cannot exclude the possibility of an additional type of selection that favors the presence of poorly translated codons in lowly expressed genes. However, we think it unnecessary to invoke such selection, because the pattern of codon usage in lowly expressed genes could simply arise from a comparatively low level of selection against nonoptimal codons.

Evidence supporting our view can be drawn from experimental results on the effect of different synonymous codons on translation rate. Several groups of researchers have obtained experimental evidence for the effect of particular codons on gene expression in *E. coli*. Insertion of the very rare codon AGG into a highly expressed gene of *E. coli* reduced the rate of translation (Robinson et al. 1984). In a different experimental system, successive replacement of three AGG codons with CGT progressively increased the level of gene expression (Bonekamp et al. 1985). However, in each case the effect was not detectable except at high rates of expression (Robinson et al. 1984; Bonekamp et al. 1985; see also the theoretical work by Varenne and Lazdunski 1986). This suggests that selection to reduce the translation rate from a moderate to a lower rate by means of "rare"-codon utilization would not be effective. Rather, selection to reduce the expression of a gene could occur by, for example, reduction of the strength of the appropriate promoter. The experiments cited above suggest that in lowly expressed genes selection against rare codons is very weak, so they can accumulate under the pressure of mutation. Further support for this hypothesis comes from examination of the rate of synonymous substitution in regulatory genes (see below).

There has been a suggestion that certain nonoptimal (rare) codons occur at extraordinarily high frequencies in regulatory genes expressed at very low levels (Konigsberg and Godson 1983). However, in the investigation that prompted the suggestion, the regulatory genes were compared with a group of 25 *E. coli* genes including many (e.g., ribosomal protein and outer membrane protein genes) that are expressed at high levels and have rather biased codon usage. We examined the frequencies of rare codons (defined by their virtual absence from very highly expressed genes; see Table 1) in each of eight regulatory genes from *E. coli* (Sharp and Li 1986). We found that these regulatory genes (which include those examined by Konigsberg and Godson 1983) do not appear to have significantly higher frequencies of nonoptimal codons than do a great number of other genes expressed at moderate or low levels

and showing low codon bias (see Table 1). Thus, the pattern of codon usage seen in regulatory genes could simply reflect a very low level of expression, and a consequent lack of selection against nonoptimal codons.

A Measure of Directional Synonymous-Codon Bias

It is desirable to quantify the degree of bias in codon usage in each gene in such a way that comparisons can be made both within and between species. One approach to this problem is to devise a measure for the degree of deviation from a postulated impartial pattern of usage, but there are difficulties in knowing the pattern of codon usage to be expected in the absence of selection. For example, indices that simply measure deviations from equal usage of synonymous codons, such as that of Lipman and Wilbur (1985) and the scaled chi square of Sharp et al. (1986), confound biases due to selection and mutation pressures.

Another approach is to assess the relative merits of different codons from the viewpoint of translational efficiency. For example, Ikemura (1985) has identified in *E. coli* and yeast certain "optimal" codons that are expected to be translated more efficiently than others, and calculated their frequencies in a gene. The "codon bias index" of Bennetzen and Hall (1982), for use with yeast genes, is similar. Such indices are certainly useful, but have several disadvantages. First, some amino acids are usually excluded because it is not clear which codons are "optimal." Second, all codons considered are classified only as optimal or nonoptimal, with no recognition that some codons within each category are better than others. For example, Ikemura (1985) treats CGU and CGC alike, as preferred codons for Arg in *E. coli,* yet the frequency of CGU is two to three times that of CGC in highly expressed genes. Third, there is no good basis for comparison between species because the proportional division of the codon table into the two categories may differ between species; for example, Ikemura (1985) identified 21 optimal codons, for 14 amino acids, in *E. coli,* and 19 optimal codons, for 13 amino acids, in yeast.

Gribskov et al. (1984) have recently proposed another index, the "codon preference statistic." This statistic is based on the ratio of the likelihood of finding a particular codon in a highly expressed gene to the likelihood of finding that codon in a random sequence with the same base composition. Gribskov et al. show that the statistic is useful for predicting the relative level of gene expression. However, the statistic has two disadvantages in the current context. First, in taking account of base composition it uses the values derived from highly expressed genes—where base composition is in fact more likely to be influenced by codon selection. Second, it is not normalized and therefore the values for two genes encoding proteins with different amino acid compositions can be quite different even if both genes use only the "best" codons. We (P.M. Sharp and W.-H. Li, submitted to *Nucleic Acids Research*) have devised a new index similar to the codon preference statistic but taking account of the above two factors. In recognition of the role of natural selection in producing high levels of codon bias, we call this statistic the Codon Adaptation Index (CAI).

We recognize that even in *E. coli* and yeast the factors determining the frequency of synonymous-codon usage are not completely understood, but we deduce that the pattern of codon usage in very highly expressed genes can reveal (1) which of the alternative synonymous codons for an amino acid is the most efficient for translation, and (2) the relative extent to which other codons are disadvantageous.

The first step, then, is to construct a reference table of RSCU values (see above) for very highly expressed genes of the organism in question. The relative adaptiveness of a codon, w_{ij}, is then the frequency of use of that codon compared with the frequency of the optimal codon for that amino acid:

$$w_{ij} = RSCU_{ij}/RSCU_{imax} \qquad (2)$$

where $RSCU_{imax}$ is the value for the most frequently used codon for the i-th amino acid. To obtain reference RSCU values, we used the 27 very highly expressed *E. coli* genes described above. A derived table of w_{ij} values is given elsewhere (P.M. Sharp and W.-H. Li, submitted to *Nucleic Acids Research*).

The CAI for a gene is then calculated as the geometric mean of the w_{ij} values corresponding to each of the codons used in that gene. That is,

$$CAI = \left(\prod_{k=1}^{L} w_k \right)^{1/L} \qquad (3)$$

where L is the number of codons (excluding AUG and UGG) and w_k is the w value for the k-th codon in the gene. Equation (3) can be more accurately computed as

$$CAI = \exp \frac{1}{L} \sum_{k=1}^{L} \ln w_k \qquad (4)$$

or, from a codon usage table,

$$CAI = \exp \frac{1}{L} \sum_{i=1}^{18} \sum_{j=1}^{n_i} X_{ij} \ln w_{ij} \qquad (5)$$

where X_{ij} and n_i are as defined in Eq. (1).

There is no intrinsic effect of gene length (L) on CAI, but CAI values for short genes may be more variable due to sampling effects.

Table 2. Comparison of genes between *Escherichia coli* and *Salmonella typhimurium*

Gene	L[a]	K_A[b]	K_S[c]	CAI[d]
trpA	267	0.083	1.773 (0.326)	0.332
cheY	128	0.010	1.492 (0.310)	0.324
trpC	451	0.071	1.391 (0.180)	0.311
tar	552	0.124	1.366 (0.138)	0.319
araC	280	0.038	1.269 (0.166)	0.242
aroA	425	0.067	1.259 (0.139)	0.332
pabA	186	0.083	1.244 (0.276)	0.288
dnaG	580	0.081	1.178 (0.107)	0.276
trpE	519	0.069	1.063 (0.099)	0.344
trpD	530	0.016	1.060 (0.103)	0.330
trpB	396	0.020	1.031 (0.114)	0.382
crp	209	0.002	0.888 (0.159)	0.472
orf1[e,f]	108	0.043	0.842 (0.173)	0.368
ilvY[e]	257	0.009	0.717 (0.095)	0.320
metB	385	0.024	0.538 (0.058)	0.332
rpoD	612	0.012	0.489 (0.045)	0.551
ilvM	86	0.032	0.468 (0.117)	0.230
glnA[e]	71	0.071	0.392 (0.120)	0.576
ompA	345	0.039	0.345 (0.046)	0.737
metJ	104	0.008	0.290 (0.079)	0.387
rpsU	70	0.000	0.039 (0.028)	0.720
5′ tar	280	0.198	1.363 (0.212)	0.305
3′ tar	235	0.012	1.113 (0.156)	0.354

Data from Sharp and Li (in press)
[a] Number of codons
[b] Number of substitutions per nonsynonymous site
[c] Number of substitutions per synonymous site. Values in parentheses are standard errors
[d] Average Codon Adaptation Index
[e] Partial sequence
[f] Open reading frame upstream of pyrE

Values of CAI clearly parallel levels of gene expression. For example, ribosomal protein genes are highly expressed and generally have high CAI values, while lowly expressed regulatory genes (e.g., lacI and trpR in *E. coli*) have low CAI values (for further discussion see P.M. Sharp and W.-H. Li, submitted to *Nucleic Acids Research*).

The CAI is a very simple measure of the extent of synonymous-codon usage bias, specifically in the direction of the bias seen in highly expressed genes. Compared with indices that measure only the frequencies of certain optimal codons, it has the advantage of taking account of all 59 codons for which synonymous alternatives exist, each in a quantitative manner. The CAI has many applications. Since it assesses the extent to which selection has molded the pattern of codon usage in a gene it is useful for predicting the level at which a gene is expressed. It is also useful for comparing codon usage biases in different organisms and for assessing the extent of adaptation of viral genes to the host translation system (P.M. Sharp and W.-H. Li, submitted to *Nucleic Acids Research*). A high CAI value indicates that an open reading frame is probably a gene, but a low value is subject to more than one interpretation (P.M. Sharp and W.-H. Li, submitted to *Nucleic Acids Research*). Below, we use the CAI to investigate the relationship between the degree of codon usage bias and the rate of synonymous substitution in a gene.

Variation in Rates of Synonymous Substitution

From comparisons of the rates of nucleotide substitution at noncoding sites (e.g., pseudogenes), coding but degenerate sites (i.e., sites at which synonymous mutations can occur), and amino acid-determining sites it is clear that the rate of molecular evolution is inversely related to the degree of selective constraint on the sequences or sites involved (Kimura 1983; Li et al. 1985a). If, as argued above, the variation among genes in the degree of codon usage bias reflects variation in the selective constraints on synonymous changes, then this should be echoed by variation among genes in the rate of synonymous substitution. More specifically, the rate of synonymous substitution should be inversely related to the degree of codon usage bias (Ikemura 1985; Kimura 1986). Analysis of a large number of mammalian genes has revealed that the synonymous-substitution rate indeed varies considerably among genes (Li et al. 1985b; W.-H. Li and M. Tanimura, unpublished data), but the patterns of codon usage in mammals are not well understood. Preliminary studies of the synonymous-substitution rate in a few enterobacterial genes (Ikemura 1985; Kimura 1986) suggest that genes with high proportions of optimal codons evolve more slowly.

We (Sharp and Li, in press) have examined DNA sequence data for 21 pairs of homologous genes from *E. coli* and *S. typhimurium* (Table 2). Divergence between genes was calculated by a method (Li et al. 1985b) that takes into account both the degree of degeneracy of nucleotide sites and the different rates of transitions and transversions to estimate the numbers of nucleotide substitutions per synonymous (K_S) and per nonsynonymous (K_A) site. The degree of synonymous-codon usage bias was measured by the CAI, using the reference set of very highly expressed *E. coli* genes to calculate the CAI values for both the *E. coli* and *S. typhimurium* genes. For each gene the CAI values for the two species are quite similar.

While a good estimate of the times of divergence of *E. coli* and *S. typhimurium* from a common ancestor does not exist, we can nevertheless compare the relative rates of evolution between genes (Table 2). A striking observation is that there is a large range of synonymous-substitution rates among genes. The two very highly expressed genes, rpsU

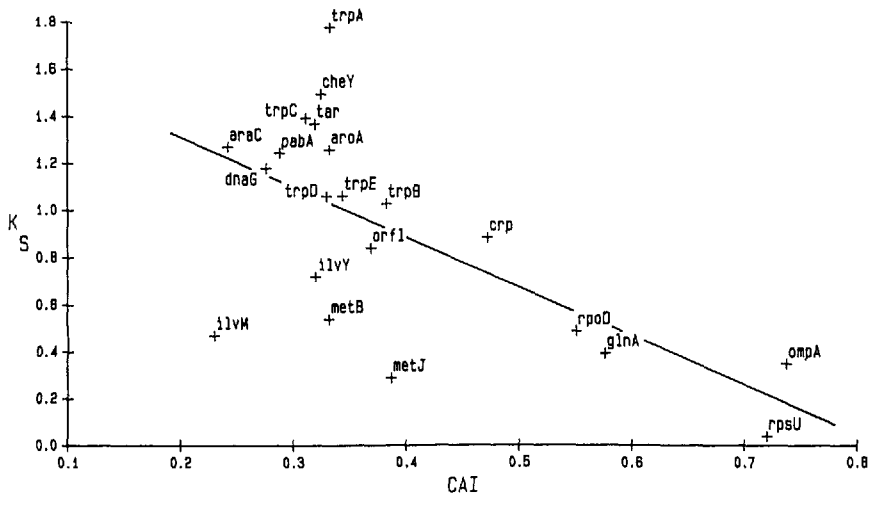

Fig. 1. Relationship between synonymous-codon usage bias and rate of synonymous substitution in *Escherichia coli* and *Salmonella typhimurium*. K_S, estimated number of synonymous nucleotide substitutions per site; CAI, Codon Adaptation Index (mean of the values for the two species). Note that for a random sequence the CAI value is 0.17. The least-squares linear regression of K_S on CAI is indicated; the linear correlation coefficient is 0.65, $P < 0.01$

and ompA, have both the greatest biases in codon usage and very low degrees of synonymous divergence (Table 2). Two genes with a high codon bias, rpoD and glnA, also show comparatively low rates of synonymous substitution. Among genes with low codon biases there is considerable variability in K_S, but high K_S values predominate among the longer sequences, which are less affected by stochastic variation.

The rate of synonymous substitution has been plotted against the degree of codon bias (Fig. 1). There is a highly significant negative correlation (linear correlation coefficient, 0.65; $P < 0.01$) between these two statistics, confirming that genes with more extreme synonymous codon biases undergo synonymous substitution at a lower rate. Horizontal transfer of genes between *E. coli* and *S. typhimurium* would produce gene pairs with surprisingly low synonymous divergence for a particular degree of codon bias. In Fig. 1 a few genes appear perhaps to be in this category, but for the two outstanding examples (metJ and ilvM) the situation can be explained by the small number of codons examined (the K_S value is subject to larger sampling errors when the number of synonymous sites is small) or by extra sequence constraints [ilvM contains within its coding sequence a promoter for the neighboring ilvE gene (Lopes and Lawther 1986)], so interspecific exchange does not appear to be an important confounding factor.

Sequence data were also examined for a few genes from *Klebsiella pneumoniae, Enterobacter aerogenes,* and *Serratia marcescens*—species of Enterobacteriaceae more distantly related to *E. coli* (Sharp and Li, in press). Again, genes with low codon usage biases have high K_S values, whereas genes with high codon usage biases (particularly lpp) show much lower synonymous divergence.

Thus the observation that molecular-evolutionary rates are inversely related to selective constraints can be extended to synonymous sites in different genes. On the other hand, the Enterobacteriaceae also follow the mammals in showing a tendency for genes with a high nonsynonymous-substitution rate to have a high synonymous-substitution rate (Li et al. 1985b). A correlation of K_A and K_S indicates that among the genes studied, those that are highly expressed and tend to have a high CAI and a low K_S also tend to encode conserved proteins. A direct relationship between protein sequence constraint and codon bias would be unlikely. This is testable because the importance of the precise amino acid sequence to protein function and hence the degree of sequence conservation (and K_A) can vary along a peptide, whereas the synonymous-codon composition is a property of an mRNA as a whole and so the degree of codon bias (and hence K_S) should be comparatively uniform along a gene. The tar gene is an example where the K_A values are very different in the 5′ and 3′ halves of the gene while the K_S values are similar (see bottom of Table 2).

As noted earlier, it has been suggested that in some cases the level of gene expression could be modulated evolutionarily by the selection of rare codons to reduce the rate of translation. In particular, it has been reported that the dnaG, lacI, trpR, and araC genes of *E. coli* have excesses of rare codons, and this finding has been explained as a mechanism to maintain low expression (Konigsberg and Godson 1983). We have noted above that these genes do not have significantly more rare codons than do a large number of other *E. coli* genes expressed at moderate to low levels. Here we point out that selection for rare codons should reduce the rate of synonymous substitution, just as selection for optimal codons in highly expressed genes reduces the rate. In Fig. 1 it can be seen that dnaG and araC are accumulating synonymous substitutions at a rate

Table 3. Mean proportion of optimal codons in a gene under a model of additive selection with a threshold[a]

u[b]	s	K[c]	Mean (SD)
10^{-5}	0.0002[d]	0	0.340 (0.033)
	0.0020	0	0.777 (0.026)
		60	0.768 (0.024)
		90	0.694 (0.006)
	0.0040	0	0.940 (0.011)
10^{-6}	0.0020	0	0.881 (0.011)
		30	0.890 (0.006)

Taken from Li (1987)

[a] A haploid population of N = 1000 is used. The number of codons (degenerate sites) in the gene is 300 (with one exception, noted below)

[b] Rate of mutation per nucleotide site per generation. At each degenerate codon site the rate of mutation from the optimal codon to a nonoptimal codon is u, while the rate of mutation from a nonoptimal codon to the optimal codon is u/3

[c] In the model of additive selection with a threshold the fitness for a gene with n nonoptimal codons is 1 if $n \leq K$ and $1 - (n - K)s$ if $n > K$. When K = 0 (simple additive selection) the fitness for a gene with n nonoptimal codons is $1 - ns$

[d] In this case the number of codons in the gene is 100

typical of genes with low codon biases. This supports the view that the incidence of rare codons in these genes results from an absence of strong negative (purifying) selection rather than the presence of positive selection.

Theoretical Models

From the population-genetic viewpoint, a very interesting question is, how much selective advantage is required to produce a certain degree of codon usage bias? Kimura (1981, 1983) seems to have been the first author to treat this problem. He uses a model of stabilizing selection, neglecting the linkage of nucleotides within a gene. Taking the relative availability of isoaccepting tRNA molecules as the major factor determining the choice of synonymous codons, he assumes that the optimal state (the highest fitness) for a gene is achieved when the relative frequencies of synonymous codons in the mRNA exactly match those of the isoaccepting tRNA species in the cell. He shows that a small selective difference among synonymous codons can produce a strong usage bias.

We (Li 1987) have recently proposed an alternative approach to the problem. We assume that the nucleotides in a gene are completely linked and that at each degenerate codon site the alternative codons for an amino acid can be categorized as either optimal (this class being denoted by B_1) or nonoptimal (B_2). There are L degenerate sites in a gene, and each B_2 site has the same selective disadvantage, s. We consider an additive selection scheme with a threshold K (where K is a nonnegative integer) such that the fitness of a gene with n B_2 sites is 1 if $n \leq K$ and $1 - (n - K)s$ if $n > K$. When K = 0 this reduces to a simple additive scheme. The fitness of a gene with n B_2 sites is then equal to $1 - ns$, and since the s values used are very small this is a good approximation to a multiplicative scheme where the fitness of the same gene is equal to $(1 - s)^n$ (Li 1987). Some results of a computer simulation of these models are shown in Table 3. These results can also be applied to other values of N (the population size), s, and u (the mutation rate) as long as the products Ns and Nu remain constant. Here, for simplicity, we consider only the situation where an amino acid is encoded by four codons, of which one is optimal.

First, let us consider the results for the simple additive scheme, i.e., when K = 0. We note that if s is one order of magnitude smaller than 1/N, then selection is ineffective and the mean proportion of B_1 sites in the gene is only 0.34, which is not far from the mean value (0.25) for the case of selective neutrality (i.e., when s = 0). When s is 1/N or larger, selection becomes effective and the proportion of B_1 sites in the gene becomes high. Thus, even a slight selective difference among synonymous codons can produce a strong codon usage bias.

Next, let us compare the results when a threshold is included. Let $\bar{q}$ ($=\bar{q}_0$) and $\bar{q}_K$ be the equilibrium mean proportions of B_1 sites per sequence, without and with a threshold, respectively. A surprising finding is that the two schemes give virtually the same results if $1 - K/L > \bar{q}$. For example, in the case of $U = 10^{-5}$ and s = 0.0020, $\bar{q}$ is 0.777 and $\bar{q}_K$ for K = 60 is 0.768. Note that in this example the selection coefficient against a sequence with 60 B_2 sites is $60 \times 0.002 = 0.12$ in the first selection scheme, but 0 in the second scheme. This result seems puzzling but can be explained as follows: If $1 - K/L > \bar{q}$, the only difference between the two schemes is that when the proportion of B_1 sites per sequence is higher than $1 - K/L$, it will on average decrease faster in the second scheme than in the first scheme. In an equilibrium population, however, the proportion of B_1 sites per sequence rarely becomes substantially higher than $\bar{q}$, and therefore the two selection schemes should be statistically almost identical.

Another interesting finding from Table 3 is that $\bar{q}_K$ is approximately equal to $1 - K/L$ if $1 - K/L < \bar{q}$. For instance, for $u = 10^{-5}$, $\bar{q}_K$ for K = 90 is 0.694, which is almost equal to $1 - K/L = 0.700$. This property can be explained as follows: Obviously, $\bar{q}_K$ cannot be higher than $1 - K/L$. On the other hand, it cannot be substantially lower than $1 - K/L$ if $1 - K/L < \bar{q}$ because when the proportion of B_1 sites per sequence is $1 - K/L$ or lower, the selective disadvantage of adding a B_2 site is s, as strong as in the first scheme, but the mutation pressure from B_1

sites to B_2 sites is weaker than when the proportion is $\bar{q}$ or higher.

To explain the phenomenon of nonrandom codon usage under the additive scheme one can assume that selection against nonoptimal codons is stronger in highly expressed genes than in moderately and weakly expressed genes, so that the proportion of optimal codons is higher in the former than in the latter. Under this scheme it is rather easy to explain the high proportion of optimal codons in highly expressed genes in *E. coli* and yeast. These two organisms have very large effective population sizes—a value of 10^6 would be a conservative estimate. Thus, a value for s of the order of 10^{-6} or 10^{-5} might be large enough to maintain a very high proportion of optimal codons in a gene. In smaller populations the s value required is larger, but even if N is as small as 10^4, it still only needs to be of the order of 10^{-4} or 10^{-3}.

One difficulty with the additive selection scheme is as follows: On the one hand, if s is as small as 10^{-5} or 10^{-6}, it is difficult to imagine how selection could operate in a population without being overwhelmed by random factors. Also, if the proportion q of optimal codons in a gene is greatly reduced by a disturbing event such as a prolonged population bottleneck, it will take a long time for q to return to the equilibrium value. This would make it difficult to explain why the proportion of optimal codons is always high in highly expressed genes in *E. coli* and yeast. On the other hand, if s is 10^{-4} or larger, the "mutational load" can be large. For instance, the selection coefficient against a gene is $30 \times 10^{-4} = 0.003$ if $s = 10^{-4}$ and the gene carries 30 nonoptimal codons. If there are 100 such genes in the genome, the reduction in fitness is 0.3, a large load. However, from the preceding results we can draw inferences about some other selection schemes where the load can be reduced to some extent.

One alternative scheme is synergistic selection, where the disadvantage of adding a B_2 site is not constant but increases with the number of B_2 sites in the sequence. Let W_i be the fitness of a sequence with i B_2 sites. Suppose that $W_i - W_{i+1}$ is initially smaller than s but becomes equal to s when $i = K$. Then from the results in Table 3, we can conclude that the equilibrium mean for the synergistic selection scheme is close to $1 - K/L$ if $1 - K/L \leq \bar{q}$, and is greater than or equal to $\bar{q}$ if $1 - K/L \geq \bar{q}$, where $\bar{q}$ is the value predicted by the simple additive scheme. This model might be realistic if the selective disadvantage of adding a nonoptimal codon to a gene is negligibly small as long as the rate of translation can meet the need of the organism, but the disadvantage increases with the total number n of nonoptimal codons in the gene if n exceeds a critical value n_c beyond which the rate of translation can no longer meet the need of the organism. For a highly expressed gene one assumes that n_c is very small and that selection against nonoptimal codons quickly becomes effective, so that a high proportion of optimal codons is maintained in the gene. For a moderately or weakly expressed gene one assumes that n_c is relatively large, so that many nonoptimal codons can be accumulated in the gene. As in the additive scheme, a small selective advantage is sufficient to maintain a high proportion of optimal codons in a gene when the population size is large.

Another alternative scheme is stabilizing selection, which assumes that the optimal state (maximum fitness) for a gene obtains when the proportion of optimal codons in a gene is equal to a certain value p. For example, suppose that the fitness W_i of a sequence is additive over sites but is maximal at $i = K > 0$ instead of $i = 0$. That is, $W_i - W_{i+1}$ is equal to s if $i \geq K$ but $-s$ if $i < K$. Under this selection scheme the equilibrium mean of the proportion of B_1 sites per sequence is the same as the value $\bar{q}$ predicted by the simple additive scheme if $1 - K/L > \bar{q}$, and is close to $1 - K/L$ if $1 - K/L \leq \bar{q}$. This is the same as Kimura's (1981, 1983) scheme, except that $p = 1 - K/L$ can be higher or lower than the proportion of the most abundant isoaccepting tRNA species in the cell. This scheme differs from the two above in that it is advantageous to use some nonoptimal codons. This might occur if the rate of translation is highest when the relative frequencies of synonymous codons match those of the isoaccepting tRNAs in the cell (Kimura 1981, 1983). However, as noted above, there is experimental evidence that in highly expressed genes the presence of nonoptimal codons (i.e., those translated by rare tRNAs) reduces the rate of translation. Also, we have argued extensively above that selection pressures on synonymous codons appear to be unidirectional, and so we consider this selection scheme less plausible than the synergistic selection scheme.

Discussion

Generality of Results

We have dealt largely with codon usage in *E. coli*, with some reference to the yeast *S. cerevisiae*. Although yeast genes generally favor codons rather different from those preferred in *E. coli*, it would appear that the patterns of synonymous-codon usage in yeast can be explained in a similar way. For example, highly expressed yeast genes show a strong bias in favor of a small subset of codons (Bennetzen and Hall 1982) that can be identified as those best recognized by the most abundant tRNA species

(Ikemura 1982). Again, codon usage in lowly expressed genes shows a lower bias, and seems to be more influenced by mutation pressure, as evidenced by a lower G + C content at synonymous codon positions (Sharp et al. 1986). Interestingly, highly expressed genes in yeasts seem to have higher codon biases than their counterparts in *E. coli* (P.M. Sharp and W.-H. Li, submitted to *Nucleic Acids Research*), and a compilation of 110 yeast genes reveals a clear difference in the degree of codon usage bias between highly and lowly expressed genes (Sharp et al. 1986). Higher bias may be due to either larger selective differences between codons or a larger effective population size.

Codon usage in *B. subtilis* is much less biased than that seen in *E. coli* or yeast (Ogasawara 1985). Nevertheless, differences in degree of bias between genes can be clearly identified and seem to be correlated with levels of gene expression (D. Shields and P.M. Sharp, unpublished data). The preferred codons are generally different from those in *E. coli* or yeast (McConnell et al. 1986). The latter point is interesting because *B. subtilis* is similar to yeast in having a low (~42%) genomic G + C content (Normore 1973).

Codon usage patterns in some viruses show signs of adaptation to the host translation system. Among coliphages, codon usage in T7 shares many features with that in *E. coli* (Sharp et al. 1985), though lambda is rather different in this regard (Grantham et al. 1985).

In multicellular organisms selective constraints may vary depending on the tissue of expression, since tRNA populations seem to vary from tissue to tissue. Hastings and Emerson (1983) could find no significant difference in codon usage between mammalian liver genes and bird muscle genes, but they examined rather few genes. As already remarked, mammalian (and bird) genomes seem, for reasons unknown, to comprise a patchwork of regions of varied G + C content (Bernardi et al. 1985) and this local genomic G + C composition seems to be the major influence on codon choice in any gene (Bernardi and Bernardi 1985; Ikemura 1985). This base-composition effect may not be strictly "mutational" in origin, in that it seems unlikely that the pattern of mutation varies greatly along the genome, but nevertheless the effect represents an influence on codon usage apparently independent of selection mediated via translation. If mutations to GC base pairs are, for some reason associated with, e.g., large-scale chromatin structure, more or less prevalent in particular chromosomal regions, then from the population-genetic perspective this is analogous to different mutational patterns in different organisms. It may be that this effect is so large as to obscure selective effects on codon usage. Perhaps more thorough investigations taking into account local G + C contents and the level and tissue of gene expression will reveal selective constraints on codon usage in mammals.

Codon usage in other organisms has not been investigated in sufficient detail. Some (generally small) compilations exist for other taxonomic groups (e.g., *Drosophila,* plants) but the pattern across different genes has not been systematically investigated.

Conclusions and Remarks

In unicellular organisms it is possible to explain codon usage patterns in terms of a balance between selection (largely mediated via translation) and mutation. The pattern of selection is probably unidirectional, so that in a particular organism the same codons are always favored, but to different extents in different genes. The observation of very high codon biases in highly expressed genes can be reconciled with very small selective differences among synonymous codons. The pattern of codon usage in a gene reflects, but does not modulate, the level of gene expression.

The rate of synonymous substitution varies among genes depending on the degree of constraint on codon choice. The differences among *E. coli* genes in rate of synonymous substitution are so large that great care must be taken when combining results from different genes in any attempt to derive a molecular clock of synonymous substitution. Although patterns of selective constraint on mammalian genes have not yet been clearly identified, the observation of different rates of synonymous substitution among mammalian genes (Li et al. 1985b; W.-H. Li and M. Tanimura, unpublished data) suggests that a more thorough analysis would be fruitful.

Acknowledgments. We thank Clay Stephens for reading the manuscript. This study was supported by NIH Grant GM30998 and NSF Grant BSR-8303965.

References

Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bibb MJ, Findlay PR, Johnson MW (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene 30:157–166

Bonekamp F, Andersen HD, Christensen T, Jensen KF (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

Bulmer M (1986) Neighbouring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322–329

Fitch WM (1976) Is there selection against wobble in codon–anticodon pairing? Science 194:1173–1774

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

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49-r62

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r79

Grantham R, Greenland T, Louail S, Mouchiroud D, Prato JL, Gouy M, Gautier C (1985) Molecular evolution of viruses as seen by nucleic acid sequence study. Bull Inst Pasteur 83: 95–148

Gribskov M, Devereux J, Burgess RR (1984) The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res 12:539–549

Grosjean H, Fiers W (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

Hastings KEM, Emerson CP Jr (1983) Codon usage in muscle genes and liver genes. J Mol Evol 19:214–218

Hinds PW, Blake RD (1985) Delineation of coding areas in DNA sequences through assignment of codon probabilities. J Biomol Struct Dyn 3:543–549

Honess RW, Bodemer W, Cameron KR, Niller H-H, Fleckenstein B, Randall RE (1986) The A+T-rich genome of *Herpesvirus saimiri* contains a highly conserved gene for thymidylate synthase. Proc Natl Acad Sci USA 83:3604–3608

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

Ikemura T (1981b) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–409

Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573–598

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

Kimura M (1968) Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral alleles. Genet Res 11:247–269

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, England

Kimura M (1986) DNA and the neutral theory. Philos Trans R Soc Lond [Biol] 312:343–354

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

Konigsberg W, Godson GN (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

Li W-H (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol, in press

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21: 58–71

Li W-H, Luo C-C, Wu C-I (1985a) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 1–94

Li W-H, Wu C-I, Luo C-C (1985b) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Lipman DJ, Wilbur WJ (1985) Interaction of silent and replacement changes in eucaryotic coding sequences. J Mol Evol 21:161–167

Lopes JM, Lawther RP (1986) Analysis and comparison of the internal promoter, pE, of the ilvGMEDA operons from *Escherichia coli* and *Salmonella typhimurium*. Nucleic Acids Res 14:2779–2798

McConnell DJ, Cantwell BA, Devine KM, Forage AJ, Laoide BM, O'Kane C, Ollington JF, Sharp PM (1986) Genetic engineering of extracellular enzyme systems of Bacilli. Ann NY Acad Sci 469:1–17

Normore WM (1973) Guanine-plus-cytosine (GC) composition of the DNA of bacteria, fungi, algae and protozoa. In: Laskin AI, Lechevalier HA (eds) Handbook of microbiology, vol 2: microbial components. CRC Press, Cleveland, pp 585–740

Nussinov R (1984) Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res 12:1749–1763

Ogasawara N (1985) Markedly unbiased codon usage in *Bacillus subtilis*. Gene 40:145–150

Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys G (1984) Codon usage can affect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Res 12:6663–6671

Schaaper RM, Danforth BN, Glickman BW (1986) Mechanisms of spontaneous mutagenesis: an analysis of the spectrum of spontaneous mutation in the *Escherichia coli* lacI gene. J Mol Biol 189:273–284

Sharp PM, Li W-H (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. Nucleic Acids Res 14:7734–7749

Sharp PM, Li W-H (1987) The rate of synonymous substitution in Enterobacterial genes inversely related to codon usage bias. Mol Biol Evol, in press

Sharp PM, Rogers MS, McConnell DJ (1985) Selection pressures on codon usage in the complete genome of bacteriophage T7. J Mol Evol 21:150–160

Sharp PM, Tuohy TMF, Mosurski KR (1986) Codon usage in yeast: Cluster analysis clearly differentiates between highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Varenne S, Lazdunski C (1986) Effect of distribution of unfavourable codons on the maximum rate of gene expression by an heterologous organism. J Theor Biol 120:99–110