Review

Codon Usage and tRNA Content in Unicellular and Multicellular Organisms[1]

Toshimichi Ikemura
Department of Biophysics, Kyoto University

Choices of synonymous codons in unicellular organisms are here reviewed, and differences in synonymous codon usages between *Escherichia coli* and the yeast *Saccharomyces cerevisiae* are attributed to differences in the actual populations of isoaccepting tRNAs. There exists a strong positive correlation between codon usage and tRNA content in both organisms, and the extent of this correlation relates to the protein production levels of individual genes. Codon-choice patterns are believed to have been well conserved during the course of evolution. Examination of silent substitutions and tRNA populations in Enterobacteriaceae revealed that the evolutionary constraint imposed by tRNA content on codon usage decelerated rather than accelerated the silent-substitution rate, at least insofar as pairs of taxonomically related organisms were examined. Codon-choice patterns of multicellular organisms are briefly reviewed, and diversity in G+C percentage at the third position of codons in vertebrate genes—as well as a possible causative factor in the production of this diversity—is discussed.

Introduction

All amino acids except Met and Trp are coded for by two to six codons. DNA-sequence data from diverse organisms clearly show that synonymous codons for any amino acid are not used with equal frequency even though choices among the codons should be equivalent in terms of protein structures. Grantham and his colleagues have found that synonymous codons are used differently by different kinds of organisms and that each type of genome has a particular coding strategy—that is, the choices among synonymous codons are consistently similar for all genes within each type of genome (Grantham 1980; Grantham et al. 1980, 1981). This finding has been designated the "genome hypothesis." It is also becoming increasingly clear that organism-specific codon choice is related to organism-specific populations of isoaccepting tRNAs, at least in the cases of *Escherichia coli* and yeast (Post et al. 1979; Post and Nomura 1980; Ikemura 1980, 1981*a*, 1981*b*, 1982; Bennetzen and Hall 1982) (isoaccepting tRNAs are tRNAs that are charged with the same amino acid but usually respond to different codons for that amino acid). I shall first explain the correlation between codon usage and isoaccepting tRNA content in these two unicellular microorganisms. Then the codon-choice patterns of other unicellular organisms will be discussed. Finally, I will briefly review the codon-choice patterns of multicellular organisms and discuss the differences in codon choice between unicellular and multicellular organisms.

1. Key words: codon, tRNA, molecular evolution, G+C percentage, DNA sequence, codon usage.

Address for correspondence and reprints: Toshimichi Ikemura, Department of Biophysics, Faculty of Science, Kyoto University, Kyoto 606, Japan.

Mol. Biol. Evol. 2(1):13–34. 1985.

0737-4038/85/0201-0005$02.00

Table 1
Codon Usage Observed for *Escherichia coli* and Yeast Genes

	E. coli					YEAST				
CODON	*tuf* A, B	r-pro.	*rpo* B, D	*thr* A, B	*trp* A–C, E	G3PDH	Enol.	Histon H2A2B	TRP5	CYC 1, 7
Leu:										
UUA ...	0	1	2	14	21	0	5	9	15	3
UUG ...	0	2	8	23	24	41[a]	73[a]	31[a]	24[a]	8[a]
CUU ...	2	4	11	10	16	0	0	0	4	1
CUC ...	1	3	18	18	16	0	0	0	4	0
CUA ...	0	0	1	3	9	1	0	4	11	1
CUG ...	53[a]	79[a]	141[a]	55[a]	96[a]	0	0	2	4	0
Arg:										
CGU ...	41[a]	48[a]	89[a]	24[a]	32[a]	0	2	0	3	0
CGC ...	5[a]	26[a]	46[a]	23[a]	51[a]	0	0	0	1	0
CGA ...	0	0	1	6	4	0	0	0	0	0
CGG ...	0	0	0	12	3	0	0	0	0	0
AGA ...	0	1	0	0	2	22[a]	26[a]	30[a]	22[a]	5[a]
AGG ...	0	0	0	2	1	0	0	2	2	1
Pro:										
CCU ...	0	3	9	6	11	1	1	2	11	4
CCC	0	0	0	6	12	0	0	0	4	0
CCA	2	4	11	5	16	22	27	18	15	5
CCG ...	38[a]	36[a]	55[a]	26[a]	40[a]	0	0	0	11	0
Gln:										
CAA	0	9	15	14	26	11	18	18	21	2
CAG ...	16[a]	33[a]	73[a]	32[a]	51[a]	0	0	2	6	3
Lys:										
AAA ...	35[a]	90[a]	77[a]	24[a]	53[a]	3	10	25	18	4
AAG ...	11	24	37	21	11	49[a]	62[a]	36[a]	26[a]	29[a]
Ala:										
GCU ...	24[a]	93[a]	30[a]	18[a]	31[a]	49[a]	96[a]	51[a]	37[a]	6[a]
GCC ...	2	10	19	44	67	16[a]	17[a]	20[a]	14[a]	6[a]
GCA ...	11[a]	45[a]	30[a]	20[a]	34[a]	0	0	1	12	2
GCG ...	17[a]	28[a]	49[a]	43[a]	60[a]	0	0	1	1	1
Val:										
GUU ...	46[a]	54[a]	55[a]	27[a]	28[a]	45[a]	32[a]	14[a]	24[a]	2[a]
GUC ...	1	6	21	22	20	27[a]	37[a]	8[a]	24[a]	1[a]
GUA ...	21[a]	40[a]	34[a]	8[a]	15[a]	0	0	0	6	1
GUG ...	6[a]	16[a]	34[a]	32[a]	44[a]	0	0	2	7	2
Gly:										
GGU ...	38[a]	49[a]	78[a]	31[a]	37[a]	49[a]	72[a]	36[a]	52[a]	6[a]
GGC ...	41[a]	34[a]	47[a]	36[a]	50[a]	0[a]	0[a]	0[a]	7[a]	3[a]
GGA ...	0	0	0	12	10	0	0	0	3	2
GGG ...	2	0	5	14	14	0	0	0	4	3
Ser:										
UCU ...	14	18	32	11	13	24[a]	28[a]	32[a]	21[a]	4[a]
UCC ...	6	18	38	17	16	26[a]	33[a]	17[a]	8[a]	1[a]
UCA ...	0	1	2	7	10	0	0	3	6	2
UCG ...	0	1	5	11	18	0	0	0	1	1
AGU ...	0	1	3	7	10	0	0	0	6	2
AGC ...	1	12	23	16	23	0	0	1	1	0
Thr:										
ACU ...	25[a]	36[a]	19[a]	5[a]	11[a]	22[a]	17[a]	18[a]	25[a]	4[a]
ACC	31[a]	26[a]	63[a]	21[a]	35[a]	25[a]	23[a]	10[a]	10[a]	4[a]
ACA	3	3	3	5	8	0	0	5	9	5
ACG ...	1	0	13	8	18	0	0	0	2	4

Table 1 (*Continued*)

Codon	*E. coli* *tuf* A, B	*E. coli* r-pro.	*E. coli* *rpo* B, D	*E. coli* *thr* A, B	*E. coli* *trp* A–C, E	Yeast G3PDH	Yeast Enol.	Yeast Histon H2A2B	Yeast TRP5	Yeast CYC 1, 7
Ile:										
AUU ...	6	13	29	34	47	16[a]	24[a]	17[a]	19[a]	5[a]
AUC ...	52[a]	51[a]	98[a]	26[a]	37[a]	23[a]	19[a]	13[a]	12[a]	3[a]
AUA ...	0	0	0	2	1	0	0	0	3	1
Asn:										
AAU ...	0	3	4	28	25	0	2	5	11	3
AAC	14[a]	42[a]	66[a]	24[a]	33[a]	26[a]	38[a]	17[a]	15[a]	[illegible]
Phe:										
UUU ...	2	10	15	16	31	0	3	2	17	5
UUC ...	26[a]	23[a]	44[a]	23[a]	30[a]	21[a]	28[a]	4[a]	15[a]	3[a]
Tyr:										
UAU ...	3	3	18	17	31	0	1	7	9	5
UAC ...	17[a]	13[a]	38[a]	11[a]	18[a]	20[a]	18[a]	9[a]	10[a]	5[a]
Glu:										
GAA ...	60[a]	61[a]	147[a]	52[a]	74[a]	29[a]	53[a]	25[a]	33[a]	[illegible]
GAG ...	13	16	46	20	31	0	0	1	9	5
Cys:										
UGU ...	2	1	5	7	10	4	2	0	6	4
UGC ...	4	6	5	16	17	0	0	0	1	1
His										
CAU ...	3	3	5	11	20	0	0	5	11	1
CAC	19	15	23	9	20	16	20	5	8	2
Asp:										
GAU ...	8	17	60	36	60	15	14	9	26	5
GAC ...	41	45	85	18	34	30	47	3	14	4

NOTE.—*tuf* = protein chain elongation factor; r-pro. = ribosome proteins (Ikemura 1982); *rpo* = RNA polymerase; *thr, trp,* and TRP = the respective amino acid synthesis genes; G3PDH = glyceraldehyde-3-phosphate dehydrogenase; Enol. = enolase; CYC1, 7 = nuclear encoding iso-1-cytochrome C and iso-2-cytochrome C, respectively. Genes with similar functions (e.g., ribosome protein genes) or those belonging to the same operon are usually treated as a collective gene, in this table as well as in tables 3 and 4 and in figure 1. References for DNA sequences and details are described by Ikemura and Ozeki (1983); see also GenBank (release 16).

[a] Optimal codons of either organism deduced by combining the predictions from rules 1–4 (see text). An example of this deduction for Arg is as follows. The most abundant *E. coli* Arg isoacceptor responds to CGU, CGC, and CGA (table 2), and therefore rule 1 predicts the preference "CGU, CGC, CGA > CGG, AGA, AGG." Because this tRNA has inosine at the anticodon, rule 3 predicts "CGU, CGC > CGA." Then the integration of these preferences is "CGU, CGC > CGA, CGG, AGA, AGG," and thus both CGU and CGC are the *E. coli* Arg optimal codons. In contrast, the most abundant yeast Arg isoacceptor responds to AGA and AGG and has 5-methoxy-carbonylmethyl U at the anticodon (rule 2 predicts "AGA > AGG"). Thus, AGA is the yeast Arg optimal codon. Such derivations for other amino acids have been extensively described by Ikemura (1981*b*, 1982) and Ikemura and Ozeki (1983). This symbol is not added to amino acids whose isoacceptors have not yet been experimentally quantified.

Organism-Specific Codon-Choice Patterns

Escherichia coli and the yeast *Saccharomyces cerevisiae* are the prokaryotic and eukaryotic unicellular organisms, respectively, whose DNA sequences have been studied most extensively. Table 1 lists examples of their codon usages. For the genes listed, as well as for most of the other sequenced genes of these organisms (~100 *E. coli* genes and ~40 yeast genes), the following characteristics of codon choices have been found. (For other genes, see Ikemura [1981*a,* 1981*b,* 1982] and Ikemura and Ozeki [1983].) (1) For most amino acids, choices among synonymous

codons are biased, and clear similarities of choice exist among the genes of each organism, in spite of the wide variety of gene functions. (2) In approximately half of the amino acids, codon choices are clearly different between the two organisms. We call this organism-specific codon choice the "dialect" of the organism. (3) The extent of the bias in codon usage found for individual genes in either organism is closely related to the level of protein production for each gene (Grantham et al. 1981; Ikemura 1981*a*, 1981*b*, 1982; Ikemura and Ozeki 1983; Bennetzen and Hall 1982; Gouy and Gautier 1982). In table 1, the genes listed on the left side of each organism's section (e.g., *E. coli tuf*AB and r-pro. and yeast G3PDH and enolase) are very highly expressed, and their bias within each dialect is extreme. These genes exclusively use one or a few synonymous codons, to the nearly complete exclusion of others. For moderately and poorly expressed genes such as the amino acid synthesis genes listed on the right side of each organism's section in table 1, the same type of dialect exists, but the extent of the bias is more moderate—that is, the codons that are exclusively used in the highly expressed genes are usually preferred, but other synonyms are also used at significant levels.

As explained in the following sections, the availability of tRNA molecules has been found to be a major factor in producing the codon dialects of *E. coli* and yeast. Codon choices observed in the foreign-type genes—such as phage, transposon, and plasmid genes—are somewhat similar to those embodied in the host dialect, but the level of similarity is clearly lower than that found among host genes (see Ikemura 1981*a*, 1981*b*, 1982). The codon-choice pattern of yeast mitochondrial genes differs totally from the pattern of its nuclear genes (Bonitz et al. 1980).

Correlation between Codon Usage and tRNA Content

Most tRNA molecules of individual species can be separated and quantified using two-dimensional polyacrylamide gel electrophoresis (Ikemura and Ozeki 1977). The contents of tRNAs in *Escherichia coli* (Ikemura 1981*a*), *Salmonella typhimurium* (Ikemura and Ozeki 1983), and *Saccharomyces cerevisiae* (Ikemura 1982), as well as the codons recognized by these tRNAs, are listed in table 2.

To explain the correlation between the frequency of use of codons and the content of the respective isoaccepting tRNA, the frequency of tRNA usage (i.e., anticodon usage) in individual genes was computed as follows. The frequency of use of a tRNA responding to a single codon was defined as the frequency of occurrence of the codon in the gene, and the frequency of a tRNA responding to multiple codons was defined as the total frequency of their occurrences (see Ikemura [1981*a*, 1981*b*] for details). Figure 1 shows examples of the correlation between synonymous codon usage and isoaccepting tRNA content for several amino acids. Both the content and the usage frequency of the most abundant isoacceptor of each amino acid are normalized at 1.0, which is indicated by a black dot (●), and the two values for other isoacceptors are plotted. Clearly, the most abundant isoacceptor is always used at the highest frequency. This is true for most *E. coli* and yeast genes thus far sequenced, but it is often not true for phage, transposon, or plasmid genes (Ikemura 1981*a*, 1981*b*, 1982). The line in each graph of figure 1 is that predicted if the frequency of isoacceptor use is proportional to its content. Data points of the highly expressed genes of *E. coli* (*tuf*AB, r-pro., and *omp*A) and yeast (G3PDH, enolase, and ADH-I) are far from this line and are projected to be as shown by the dashed lines. This means that the dependence of isoacceptor usage on its content is

Table 2
Codon Recognition and Relative Content of tRNAs in *Escherichia coli*, *Salmonella typhimurium*, and Yeast

E. coli AND *S. typhimurium*				YEAST		
		tRNA Content				
tRNA	Codon	*E. coli*	*S. typhimurium*	tRNA	Codon	tRNA content
Leu:				Leu:		
1	CUG	1.0	1.0	3	UUG	1.0
2	CUU, CUC	0.3	0.2	1	CUA, CUG	0.47
UUR	UUA, UUG	0.25	0.2			
CUA	CUA	minor	minor			
Val:				Val:		
1	GUA, GUG, GUU	1.05	0.9	1	GUU, GUC, GUA	1.04
2	GUC, GUU	0.4	0.2	2b	GUG	0.15
Gly:				Gly	GGU, GGC	1.41
3	GGU, GGC	1.1	0.9			
2	GGA, GGG	0.15	0.2			
1	GGG	0.1				
Ala 1	GCU, GCA, GCG	1.0	1.0	Ala 1	GCU, GCC, GCA	0.93
Arg:				Arg:		
1, 2	CGU, CGC, CGA	0.9	0.7	3	AGA, AGG	0.90
CGG	CGG	minor		2	CGU, CGC, CGA	0.28
AGR	AGA, AGG	minor		AGG	AGG	0.09
Ile:						
1	AUU, AUC	1.0	1.0			
2	AUA	0.05				
Lys	AAA, AAG	1.0	0.9	Lys:		
				2	AAG	0.76
				1	AAA, AAG	0.36
Glu 2	GAA, GAG	0.9	0.9	Glu 3	GAA, GAG	0.96
Asp 1	GAU, GAC	0.8	1.0	Asp	GAU, GAC	1.32
Thr 1 + 3	ACU, ACC	0.8	0.6	Thr 1	ACU, ACC, ACA	0.94
Asn	AAU, AAC	0.6	0.5			
Gln:						
2	CAG	0.4	0.4			
1	CAA	0.3	0.3			
Tyr 1 + 2	UAU, UAC	0.5		Tyr	UAU, UAC	0.88
Ser:				Ser:		
3	AGU, AGC	0.25	0.2	2	UCU, UCC, UCA	1.17
1	UCU, UCA, UCG	0.25	0.3	UCR	UCA, UCG	0.38
UCG	UCG	minor				
His	CAU, CAC	0.4	0.3	His	CAU, CAC	0.71
Trp	UGG	0.3	0.2	Trp	UGG	0.69
Pro:						
3	CCG, CCA, CCU	major	major			
2	CCC	minor	minor			
1	CCG	major	major			
Phe	UUU, UUC	0.35	0.2	Phe	UUU, UUC	0.76
Cys	UGU, UGC	minor		Cys	UGU, UGC	0.39
Met_m	AUG	0.3	0.4	Met_i	AUG	0.31

NOTE.—Relative abundance of individual tRNA species was measured on the basis of molecular numbers in cells, which were determined by two-dimensional polyacrylamide gel electrophoresis (Ikemura 1981*a*, 1982). The amount of Leu 1 tRNA of *E. coli* or *S. typhimurium* is normalized at 1.0, and this value was estimated to be on the order of 10^4 molecules per cell for normally growing *E. coli* (Ikemura 1981*a*). Yeast Leu 3 is normalized at 1.0. Recent data on tRNA contents are included.

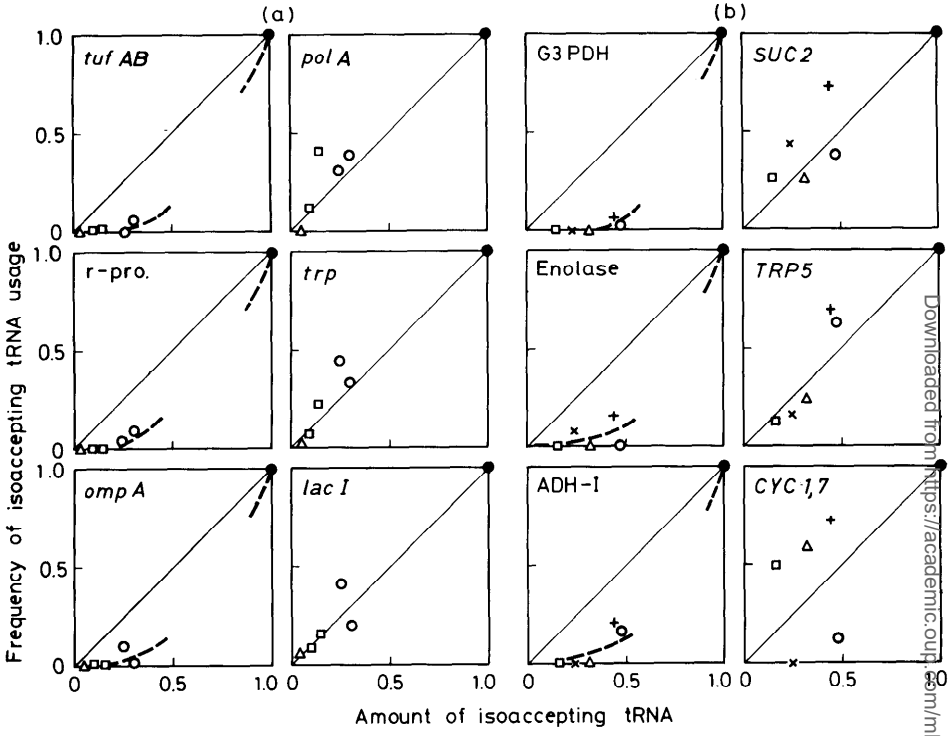

FIG. 1.—The relationship between the content of isoaccepting tRNA and the frequency of its usage for individual genes. Both the amount and the frequency of use of the most abundant tRNA of each amino acid are normalized at 1.0. The straight line in each graph is that predicted if the use of isoaccepting tRNA is proportional to its content. The most abundant isoacceptor of each amino acid is indicated by a black dot (●). Other isoacceptors are (*a*) *Escherichia coli* Leu (○), Gly (□), and Ile (△); and (*b*) yeast Arg (×), Leu (○), Ser (△), Val (□), and Lys (+). Details as well as results for other genes have been presented by Ikemura (1980, 1981*a*, 1981*b*, 1982). *omp*A = major outer membrane protein II; *pol*A = DNA polymerase I; *lac*I = *lac* operon repressor; ADH = alcohol dehydrogenase; SUC = invertase.

much greater than that which would be expected from a direct proportionality (Ikemura 1980, 1981*a*, 1981*b*). In other words, these genes selectively use codons recognized by the most abundant isoacceptors but almost completely avoid using codons of other isoacceptors; that is, the degree of their codon bias is extreme. The data points in moderately or poorly expressed genes (*E. coli pol*A, *trp*, and *lac*I and yeast SUC2, TRP5, and CYC1, 7) seem to be represented roughly by a linear function; that is, the degree of their bias is moderate. We have concluded, therefore, that codon choices in both *E. coli* and yeast genes are constrained by tRNA availability (a property that has been called "Rule 1" by Ikemura and Ozeki [1983]) and that this constraint is especially evident for highly expressed genes (Ikemura 1981*a*, 1981*b*). It should be mentioned that the synonymous codon to which the most abundant isoacceptor responds often differs between *E. coli* and yeast; for example, the most abundant Leu isoacceptor of *E. coli* responds to CUG but that of yeast responds to UUG (table 2). This is one causative factor in the production of the organism-specific codon-choice pattern.

Preference Among Codons Recognized by a Single tRNA

A single tRNA can usually respond to multiple codons (see table 2), as proposed by Crick's (1966) "wobble hypothesis." The following definite constraints

on choices from among the codons recognized by a single tRNA have been disclosed for both *Escherichia coli* and yeast genes (Ikemura 1981*a*, 1981*b*, 1982). (1) Thiolated uridine or 5-carboxymethyl uridine at the anticodon wobble position produces, in vitro, a preference for an A-terminated codon over a G-terminated codon (Nishimura 1978; Weissenbach and Dirheimer 1978). This preference ("Rule 2") has been demonstrated for the respective codon choices in both *E. coli* and yeast genes (Ikemura 1981*a*, 1981*b*, 1982). (2) Inosine at the anticodon wobble position produces a preference for U- and C-terminated codons over the A-terminated codon ("Rule 3"), presumably to avoid unorthodox purine-purine pairing (Ikemura 1981*b*, 1982; Bennetzen and Hall 1982). (3) A codon of the (A/U)-(A/U)-pyrimidine type, which has an intrinsically weak interaction with an anticodon at the first and second positions, would support an optimal interaction strength between codon and anticodon if the third letter of the codon is C ("Rule 4") (Grosjean *et al.* 1978; Grosjean and Fiers 1982).

Table 3 shows codon preferences in *E. coli* genes predicted by these rules. Not only the listed genes but also most *E. coli* and yeast genes thus far sequenced, especially the highly expressed genes, conform very well to the expectations derived from these rules (for analyses of the preferences of other *E. coli* and yeast genes, see Ikemura [1981*a*, 1981*b*, 1982]; Ikemura and Ozeki [1983]). However, this is often not true for poorly expressed genes or for phage, transposon, and plasmid genes (Ikemura 1981*a*, 1981*b*, 1982). It should be noted here that tRNA species that have these modified nucleotides at the anticodon wobble position often differ between *E. coli* and yeast. This difference at the anticodon wobble position is also a causative factor in production of the organism-specific codon-choice pattern.

Combining Rules 2–4 with Rule 1 (i.e., the constraint due to tRNA content), the synonymous codon preferences of most amino acids are predicted for both *E. coli* and yeast. The order of preference thus hypothesized has been extensively described (Ikemura [1981*b*, 1982]; Ikemura and Ozeki [1983]; see the legend of table 1 for the case of Arg). The codons predicted to be the most preferred for individual amino acids based on these rules are thought to be the codons that are optimal for the translation system of the organism, and they are designated "optimal codons" (Ikemura 1981*b*). In table 1, the optimal codons are marked by superscript

Table 3
Preferential Usage Observed Between Codons Recognized by a Single *Escherichia coli* tRNA

tRNA	Rule	Codon Variation	Use of Codons in Individual Genes				
			*tuf*AB	r-pro.	*rpo*BD	*unc*	*trp*
Lys	2	AAA/AAG	35/11	90/24	77/37	52/19	53/11
Glu 2	2	GAA/GAG	60/13	61/16	147/46	60/22	74/31
Arg 1, 2 ...	3	CGY/CGA	46/0	74/0	135/1	60/3	83/4
Phe	4	UUC/UUU	26/2	23/10	44/15	21/10	30/31[a]
Ile 1	4	AUC/AUU	52/6	51/13	98/29	57/25	37/47[a]
Tyr	4	UAC/UAU	17/3	13/3	38/18	22/7	18/31[a]
Asn	4	AAC/AAU	14/0	42/3	66/4	31/9	33/25

NOTE.—Y = pyrimidine; *unc* = membrane-bound ATPase. All cases that can be judged by Rules 2–4 are presented. For example, among *E. coli* tRNAs, only the tRNA Arg 1, 2 has inosine at the anticodon wobble position, and thus rule 3 was applied only to this tRNA. Details have been described by Ikemura and Ozeki (1983).

[a] Indicated rule not satisfied.

"a." The spectrum of optimal codons corresponds strikingly well to the spectrum of the preferred codons of each organism (Ikemura 1982). This shows that the synonymous codon choices in these two organisms are restricted by the common constraints related to the efficiency of the translation process. Thus, the differences in the spectrum of optimal codons (i.e., the differences in dialect) between the two organisms are mainly attributable to differences in the actual population of isoacceptors and in the modified nucleotides at the anticodon wobble positions.

It should be noted here that we proposed the above rules at a time when only few *E. coli* and yeast protein genes had been sequenced (Ikemura 1980, 1981*a*, 1981*b*). Now, more than 100 *E. coli* genes and approximately 40 yeast genes have been sequenced, and the codon-choice patterns of most of them are well explained within the framework of the proposed rules. This indicates that codon-choice patterns of most genes of either genome—at least the patterns of those with high and moderate expressivity—can be explained by these common rules. The codon choices in genes of low expressivity will be discussed later from an evolutionary viewpoint.

Codon-Choice Pattern and Gene Expressivity

Figure 2 shows the distribution of optimal (o) and nonoptimal (X) codons in *Escherichia coli* and yeast genes; for Met and Trp, a single codon corresponds to each amino acid and is indicated separately by a dash. Highly expressed genes (ribosome protein genes) mostly use optimal codons, showing that their codon choices are determined by the constraints imposed by tRNA availability and the efficiency of codon-anticodon pairing. In moderately expressed genes such as the amino acid synthesis genes (*E. coli trp*A and yeast TRP5), optimal codon usage clearly decreases. To examine this tendency quantitatively, the frequency of use of optimal codons (F_{op}) was defined as the number of o's divided by the sum of the number of o's and X's.

We previously reported the F_{op} value for almost all *E. coli* and yeast genes that had thus far been sequenced (Ikemura and Ozeki 1983) and pointed out that highly expressed genes always have high F_{op} values and that these values decrease with reduced levels of protein production. Table 4 lists the F_{op} values of the *E. coli* genes whose protein contents are known from the data of Neidhardt and his colleagues (Pedersen et al. 1978). These authors have quantified the cellular contents of many *E. coli* protein molecules using two-dimensional gel electrophoresis. Since protein molecules are usually metabolically stable, their cellular contents correspond well to levels of gene expression. Figure 3, in which these two values (i.e., F_{op} values and protein contents) are plotted, reveals their close correlation (Ikemura 1981*b*). This is quantitative and conclusive evidence for the correlation between gene expressivity and the codon-choice pattern (i.e., the extent of the codon bias). The same kind of correlation has also been pointed out by other research groups, both for *E. coli* (Grantham et al. 1981; Gouy and Gautier 1982; Grosjean and Fiers 1982) and for yeast (Bennetzen and Hall 1982). The evolutionary significance of this correlation will be discussed in a later section.

Codon Choices in Genes of Other Unicellular Organisms

In the preceding sections, essentially the same conclusion concerning codon choices was deduced for a prokaryote (*Escherichia coli*) and a eukaryote (yeast)—regardless of differences in the actual codon dialects they used. This conclusion should therefore be true for a wide range of organisms, at least for a wide range of

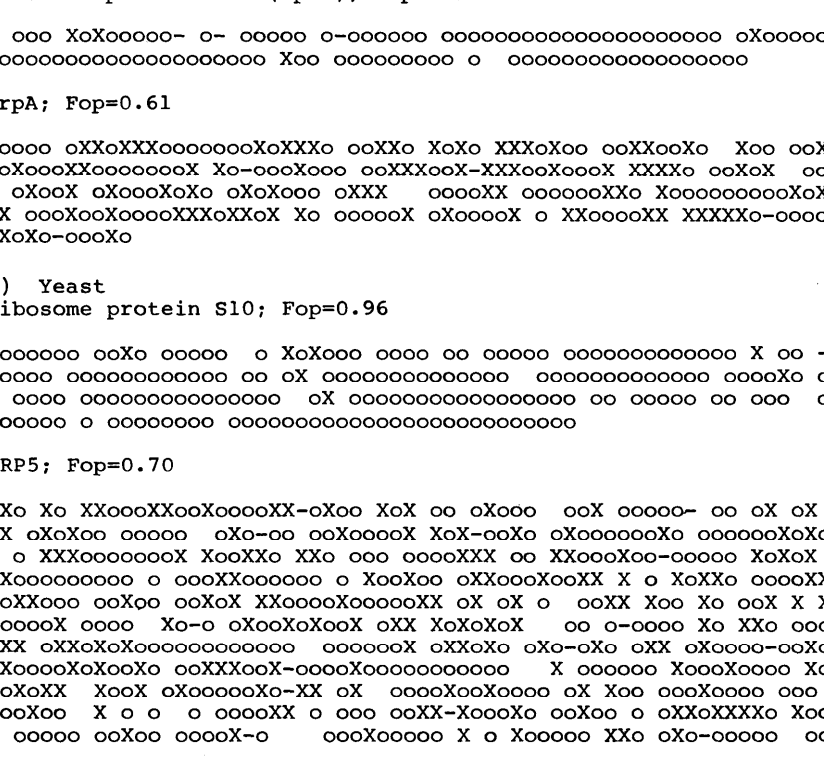

FIG. 2.—Occurrences of optimal (o) and nonoptimal (X) codons in *Escherichia coli* and yeast genes. Codons of Met and Trp are indicated by a dash (–). Blank spaces are used for amino acids, either because the contents of their isoacceptors have not been clarified or because no criteria are deduced (see Ikemura 1982). Results for other genes have been presented previously (Ikemura 1981*b*, 1982; Ikemura and Ozeki 1983).

unicellular organisms. Extensive examination of the codon choices of various prokaryotes, has revealed that the dialect of *E. coli* is similar to those of other Enterobacteriaceae (e.g., *Shigella, Salmonella, Klebsiella, Serratia,* and *Erwinia*) but differs from that of taxonomically distant organisms (e.g., *Anabaena* and *Bacillus*) (Ikemura 1982). This correlation between taxonomic distance and the extent of resemblance of codon dialect indicates that codon dialects have been conserved fairly well during evolution. Table 2, as well as our previous work (Ikemura and Ozeki 1983), shows that the tRNA population of *Salmonella typhimurium* is very similar to that of *E. coli.* Recently, we have quantified *Serratia marcescens* tRNAs. Its tRNA population also resembles *E. coli*'s, at least from a qualitative point of view (unpublished data). These findings indicate that the population of tRNA molecules has also been well conserved during evolution. Yanofsky and his colleagues (Nichols et al. 1981; Yanofsky and vanCleemput 1982) have shown that codon choice in organisms with a high G+C percentage in the genome (e.g., *S. marcescens*) is affected by this particular base composition. Codon choices of organisms whose genomic G+C content is extreme seem to be determined by the combined constraints imposed by the genomic G+C content and the tRNA content (Ikemura and Ozeki 1983).

Table 4
Fraction of Optimal Codons per Gene (F_{op}) and Number of Corresponding Protein Molecules per *Escherichia coli* Genome

Gene	F_{op}	No. of molecules
lpp	0.98	1.5×10^5
*rpl*L	0.96	4.5×10^4
*tuf*A	0.93	4.5×10^4
*omp*A ...	0.92	3×10^4
r-pro. ...	0.90	1.5×10^4
*unc*A ...	0.87	5×10^3
lpd	0.85	6×10^3
*unc*D ...	0.85	5×10^3
*rpo*B	0.83	2×10^3
*gly*Sb ...	0.80	1.5×10^3
*gln*S	0.78	1×10^3
*thr*S	0.71	1×10^3

NOTE.—*lpp* = murein lipoprotein; *rpl*L = the ribosomal protein L7/L12 present at 4 moles/ribosome; *lpd* = lipoamide dehydrogenase; *gly*Sb, *gln*S, and *thr*S = the respective aminoacyl tRNA synthetases. Numbers of *E. coli* protein molecules per genome are mainly according to Pedersen et al. (1978). For details, see Ikemura (1981*b*) and Ikemura and Ozeki (1983).

DNA-sequence data of unicellular eukaryotes other than *Saccharomyces cerevisiae* are rather restricted, and I intend to discuss only the case of *Neurospora crassa,* for which several genes with differing functions have been sequenced.

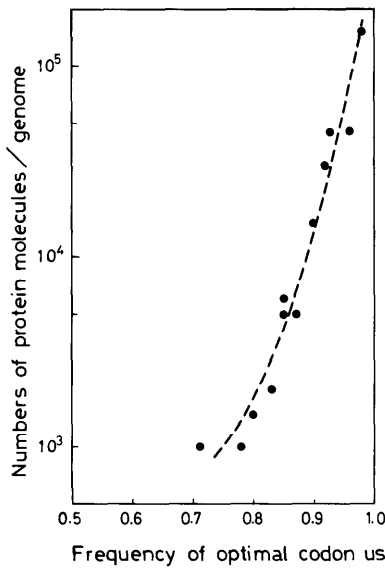

FIG. 3.—Relationship between the frequency of optimal codon use (F_{op}) and the number of *Escherichia coli* protein molecules per genome. The data of table 4 are graphed. The F_{op} value for an idealized DNA sequence in which all 61 codons encoding amino acids are evenly used is 0.47, and this may be considered a basal level frequency when actual genes are examined.

Regardless of differences in function, these genes have similar codon-choice patterns, as has been found to obtain in *S. cerevisiae.* However, the actual codon dialect differs from that of *S. cerevisiae* for Leu (CUC being most preferred by *N. crassa* and UUG the most preferred by *S. cerevisiae*), Arg(CGC and AGA, respectively), Pro(CCC and CCA, respectively), Gln(CAG and CAA, respectively), Glu(GAG and GAA, respectively), and Gly(GGC and GGU, respectively). The genomic G+C percentage of *N. crassa* is ~15% higher than that of *S. cerevisiae,* and the difference in their codon dialects for these six amino acids appears to be due at least in part to the difference in this genomic G+C content.

Evolutionary Viewpoints

The close correlation between codon usage and isoaccepting tRNA population has resulted from the accumulation of a great number of mutations and successive base substitutions occurring in both protein and tRNA genes during the course of evolution. The evident similarities of synonymous codon-choice patterns among the different genes within each unicellular organism is strong supporting evidence for the selective constraint imposed by the isoaccepting tRNA content on codon choice, although adjustment of tRNA content to the codon spectrum may also have been an important factor in establishing the correlation (Ikemura 1981*a,* 1981*b*). It should be noted that if only the adjustment of isoacceptor content to the codon spectrum is considered, one must postulate an additional and unknown mechanism by which codon-choice patterns are roughly equalized among the different genes of one organism.

Let us discuss the molecular mechanism by which synonymous codon choice has been constrained during evolution by tRNA availability. The cellular process of protein synthesis is known to require a large amount of energy and mass. In the case of the *Escherichia coli* cell, ~70% of the total cellular energy is used in this process, and the mass of ribosomes approaches one third of the dry mass of the cell. We have proposed that a codon dialect should be a reflection of the organism's strategy for producing high amounts of protein with minimal load (Ikemura and Ozeki 1983); Gouy and Grantham (1980) have extensively analyzed the dynamic aspects of *E. coli* protein synthesis (e.g., tRNA cycling) on the basis of quantitative data. The level of protein production is believed to be closely related to the level of mRNA production (e.g., Bennetzen and Hall 1982). If codons translated by minor isoacceptors were frequently used in a highly expressed gene, ribosomes would perform the uneconomical task of finding the proper tRNA present in small quantities for the large number of mRNA molecules of this gene. Furthermore, energy (GTP) seems to be consumed during this period for the purpose of "proofreading"; it has been proposed that a certain level of GTP hydrolysis is involved in rejecting tRNAs incorrectly bound to ribosomes, especially near-cognate tRNAs whose anticodon sequences are similar to and confused with that of the proper tRNA (Hopfield 1974; Thompson and Stone 1977; Thompson et al. 1981). If a mutation from an optimal codon (responded to by a major tRNA) to a nonoptimal codon (responded to by a minor tRNA) occurs, the collision frequency of ribosomes with incorrect tRNAs at the respective codon position increases, and therefore the level of GTP hydrolysis presumably also increases. The resultant loss of productive ribosome working time and of GTP energy will bring on phenotypic effects such as a decrease in growth rate and/or viability. This should become especially important as the protein production level (and therefore the number of

mRNA molecules) increases. To explain this view evolutionarily, the following equation has been proposed (Ikemura 1981*a*, 1981*b*):

$$\Delta w = P_n \times \Delta r, \quad (1)$$

where Δw is the small fitness change that would be introduced by the change in the rate of protein synthesis per mRNA (Δr) caused by a synonymous mutation. P_n is a parameter characteristic of the individual genes meant to show that in different genes the same synonymous change will produce different levels of the fitness change. Synonymous mutations occurring in a gene with a high P_n value will have a greater effect on fitness than those occurring in a gene with a low P_n value, so codon choice in the former would be more highly constrained. The P_n values are closely connected to the levels of protein production of individual genes (Ikemura 1981*a*, 1981*b*).

Next we will discuss the correlation between the extent of codon bias and the protein production level of the gene. The highly expressed genes use mostly optimal codons, showing that their codon choices are approximately optimal for translation efficiency. As the protein production (and mRNA production) level decreases, so does the optimization level (fig. 3). I have proposed that this is attributable in part to a randomization effect caused by mutation and have formulated it as follows (Ikemura 1981*b*). If the absolute value of the fitness change ($|\Delta w|$) is below a certain level (a), the mutation can be regarded as selectively neutral (Kimura 1968; King and Jukes 1969), even when the mutation produces a change in translation rate (Δr); i.e., $P_n \times |\Delta r| \leq a$. The proportion of mutations that can be regarded as neutral or nearly neutral should be larger for genes with low P_n values (low gene expressivity) than for genes with high P_n values (high gene expressivity). The present DNA sequence may represent an equilibrium or a near equilibrium state between the selective force and the random drift of neutral mutations, and F_{op} values (i.e., the extent of the codon bias) may serve as indices for this balancing point. This evolutionary view will be further explained in the following section in an examination of silent substitution rates. The correlation between the extent of the codon bias and gene expressivity has been also discussed from several other points of view (Ikemura 1981*a*, 1981*b*; Bennetzen and Hall 1982; Gouy and Gautier 1982).

Table 2 shows that in *Salmonella typhimurium*, the tRNA population is essentially the same as that in *E. coli*. We can therefore examine the effect of the common constraint imposed by this tRNA population on the rate of silent substitution (nucleotide substitution that does not cause amino acid replacement) occurring between these organisms, that is, examine whether this constraint accelerates or decelerates the silent substitution rate. Protein genes that have been sequenced for both organisms are confined at present to the tryptophan operon genes, *trp*A and B, *ara*C, and *omp*A (fig. 4A). (The *trp*C, D, and E genes were omitted because the results are essentially the same as those for *trp*A and B.) Amino acid positions where the same codon is used are indicated by double dashes (=) and those of silent substitution are indicated by a number sign (#). No marks have been added for positions of amino acid replacement. Clearly, a major portion of the nucleotide substitutions are silent substitutions (Yanofsky and vanCleemput 1982). Since we examine the effect of the constraint imposed by tRNA availability (as well as that of codon-anticodon binding) on the rate of silent substitutions, amino acid replacements were omitted from the calculations. The fraction of silent substitution was

defined as the number of #'s divided by the sum of the number of #'s and ='s; the resultant values are listed in figure 4. For the purpose of studying the silent substitution rate, it is necessary to divide silent substitutions into three groups—that is, one-base, two-base, and (for Ser only) three-base substitutions—and to correct the above values for this factor. The values as thus corrected are listed in parentheses, but this correction turned out to produce only minor changes. The levels of silent substitutions in moderately or poorly expressed genes (*trp* genes and *ara*C) are clearly higher than that of the highly expressed gene (*omp*A encoding a major membrane protein), and the substitutions in the former genes almost reached a saturation level. Figures 4B and 4C give examples for other Enterobacteriaceae. In the pair of *E. coli* and *Serratia marcescens,* whose tRNA populations roughly

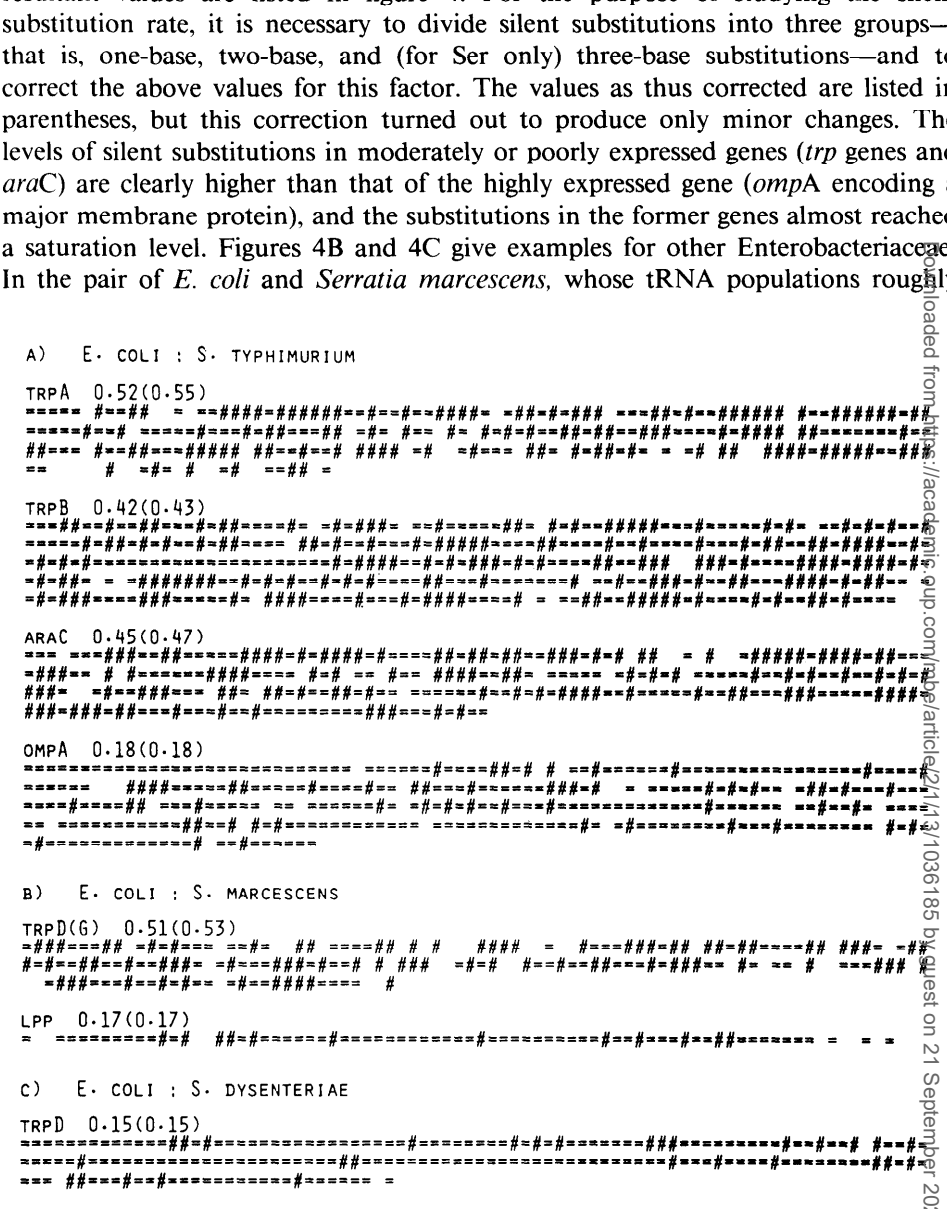

FIG. 4.—Silent substitutions observed for pairs of Enterobacteriaceae. An alignment for each pair of homologous gene sequences was done at the protein level because a large portion of the amino acid sequence is well conserved between the examined organisms. Amino acid positions where the same codon is used are indicated by the double dashes (=) and those of silent substitutions by crosshatches (#). Blanks indicate positions of amino acid replacements as well as a few gaps (deletions or insertions). *ara*C = regulatory gene for arabinose operon.

Table 5
Codon Usage Observed for Genes of Multicellular Organisms

	Number of Codons Used in								
						Individual Human Genes			
Codon	Human	Rat	Chicken	Fish	Plant	Glob. Z	Act. S	Factor 9	Feto. A
Leu									
UUA ...	61	25	3	3	24	0	0	6[a]	10
UUG ...	126	74	25	12	82[a]	1	0	3	11[a]
CUU ...	108	92	20	11	92[b]	0	2	9[b]	9
CUC ...	203[a]	127[a]	74[a]	33[a]	74	3[a]	6[a]	5	4
CUA ...	62	36	8	6	52	1	0	2	0
CUG ...	496[b]	319[b]	185[b]	59[b]	60	12[b]	18[b]	3	17[a]
Ser									
UCU ...	162	89	37	15	55	2	0	5	8
UCC ...	207[a]	143[b]	85[b]	29[b]	59[b]	7[b]	15[b]	5	11
UCA ...	99	40	12	6	45	0	0	6[a]	10[a]
UCG ...	51	19	37	7	16	1	4[a]	0	1
AGU ...	99	55	13	8	38	0	0	7[b]	2
AGC ...	212[b]	102[a]	75[a]	17[a]	58[a]	3[a]	4[a]	3	9[a]
Arg									
CGU ...	38	48[a]	35	12	23	0	1	1	1
CGC ...	113	82[b]	81[b]	22[a]	22	4[b]	12[b]	1	2
CGA ...	53	25	9	8	10	0	1	6[a]	0
CGG ...	64	37	23	11	14	0	0	3	0
AGA ...	127[b]	46	15	30[b]	37[b]	0	1	7[b]	12[a]
AGG ...	115[a]	45	40[a]	18	31[a]	2[a]	3[a]	1	5
Val									
GUU ...	118	77	33	9	75	0	0	22[b]	11
GUC ...	174	98	91	20	58	4	6	3	6
GUA ...	57	37	11	6	33	1	0	5	6
GUG ...	329[b]	187[b]	159[b]	28[b]	83[b]	6[b]	15[b]	7	16
Pro									
CCU ...	174	114	33	14	57	0	4	5	9
CCC ...	183[b]	124[b]	104[b]	23[b]	36	1	9[b]	3	4
CCA ...	94	90	20	15	80[b]	0	1	9[b]	8
CCG ...	49	21	20	6	16	4[b]	5	0	0
Thr									
ACU ...	170	87	37	13	52	2	1	12[b]	10
ACC ...	257[b]	180[b]	126[b]	24[b]	65[b]	8[b]	22[b]	7	12
ACA ...	126	80	27	7	31	0	1	10	11
ACG ...	53	30	32	6	20	2	3	1	1
Ala									
GCU ...	259	161	115	29	100	0	4	10[b]	15
GCC ...	333[b]	204[b]	199[b]	100[b]	105[b]	12[b]	19[b]	5	11
GCA ...	140	83	48	39	84	0	1	9	2
GCG ...	71	27	56	9	44	4	5	0	0
Gly									
GGU ...	108	97	50	17	62	0	4	8	4
GGC ...	279[b]	150[b]	155[b]	21[b]	76[b]	5[b]	20[b]	9	5
GGA ...	169	83	36	17	56	0	0	14[b]	13
GGG ...	125	66	58	7	27	1	4	4	4
Ile									
AUU ...	136	114	38	4	77	1	2	17[b]	15[b]
AUC ...	265[b]	200[b]	158[b]	20[b]	95[b]	6[b]	28[b]	7	8
AUA ...	62	34	6	4	27	0	0	1	11
Phe									
UUU ...	197	98	40	13	59	0	1	12[b]	17[b]
UUC ...	277[b]	154[b]	98[b]	32[b]	104[b]	7[b]	11[b]	9	15
Tyr									
UAU ...	118	91	23	4	41	0	2	11[b]	9[b]
UAC ...	194[b]	125[b]	82[b]	14[b]	81[b]	3[b]	14[b]	5	8

Table 5 (*Continued*)

	NUMBER OF CODONS USED IN								
						Individual Human Genes			
CODON	Human	Rat	Chicken	Fish	Plant	Glob. Z	Act. S	Factor 9	Feto. A
His									
CAU ...	111	44	30	5	29	0	0	6[b]	13[b]
CAC	142[b]	98[b]	79[b]	12[b]	40[b]	7[b]	9[b]	4	3
Gln									
CAA	123	80	15	7	161[b]	0	0	7	23[a]
CAG ...	315[b]	251[b]	100[b]	36[b]	83[b]	3[b]	11[b]	7	17
Asn									
AAU ...	159	90	32	6	58	0	2	15	10
AAC	253[b]	153[b]	119[b]	37[b]	113[b]	1[b]	10[b]	17[b]	10
Lys									
AAA ...	257	127	60	23	60	0	5	12	28[a]
AAG ...	413[b]	237[b]	290[b]	27[b]	115[b]	9[b]	14[b]	17[b]	14
Asp									
GAU ...	251	147	59	14	86[b]	0	4	12[b]	15
GAC ...	303[b]	201[b]	122[b]	49[b]	75	8[b]	18[b]	7	8
Glu									
GAA ...	326	173	58	18	84	0	3	33[b]	40[a]
GAG ...	434[b]	281[b]	172[b]	62[b]	112[b]	6[b]	25[b]	10	19
Cys									
UGU ...	132	111	21	13	6	0	1	19[b]	13
UGC ...	189[b]	140[b]	51[b]	19[b]	21[b]	1[b]	5[b]	5	19[a]

NOTE.—Glob. Z = zeta globin; Act. S = skeletal actin; Feto. A = alpha-fetoprotein. Except for partially sequenced genes or multigene families, codon usages in almost all genes compiled by GenBank (release 16) were summed up for the individual organisms; examples of the summed genes are presented in table 6.

[a] Second most preferred codon in a six codon box.

[b] Most preferred codon in each amino acid.

resemble each other, the highly expressed gene (*lpp* encoding a major membrane protein) again has a lower substitution rate than the *trp* gene. *Shigella dysenteriae* is a species very closely related to *E. coli*. In this pair, even the *trp* gene is far from a saturation level of substitution, and the highly expressed gene (*omp*A) again has a much lower substitution level. All the results so far available point out that the rate of silent substitution decreases as gene expressivity increases. Combining these findings, I propose that the constraint due to tRNA availability decelerates, rather than accelerates, the rate of silent substitution as far as pairs of taxonomically related organisms are concerned. This is consistent with the evolutionary view described above and with the prediction from the neutral theory (Kimura 1968, 1981, 1983). Applying the concept of stabilizing selection, Kimura has shown, on theoretical grounds, that (1) the rate of nucleotide substitution is slowed down rather than accelerated by the constraints that have produced nonrandom codon choice and (2) nonrandom choice is explained within the framework of the neutral theory (Kimura 1981, 1983). When silent substitutions occurring between taxonomically distant organisms are considered, the change in the composition of tRNA populations during evolution should be taken into account. The effects of this change (presumably caused by the change in tRNA gene copy numbers) on codon usages have been discussed (Ikemura 1981*b*). Gouy and Gautier (1982) have discussed nonrandom codon choice, stressing the overall optimization of gene sequences within a genome.

Codon Choices of Multicellular Eukaryotes

In this and the successive sections, I intend to briefly review the codon choices of multicellular eukaryotes and discuss the differences between multicellular and unicellular organisms. The DNA sequence data were taken from the GenBank (released March and November, 1983) of Los Alamos National Laboratory (Kanehisa 1982). Table 5 sums up the codon usages of most genes so far determined for man (39 genes), rat (28 genes), chicken (20 genes), fishes (trout, carp, anglerfish, catfish, eel, and salmon; 12 genes) and higher plants (pea, wheat, barley, soybean, and maize; 13 genes). To avoid an artificially excessive contribution of multigene families, only a few typical examples in each family were used for the summation. Codon usages summed for each of the vertebrates turned out to converge in an essentially identical pattern, with a possible exception of Arg; this pattern was different from the pattern of higher plants and from the dialect of *E. coli* or yeast. In this vertebrate pattern, the most preferred codon in almost all amino acids is either C- or G-terminated, but this is often not true for higher plants. The richness of G+C bases at the third position of codons of vertebrate genes cannot be a consequence of their genome-base composition, since their overall genome G+C percentage is known to range from ~40% to ~45%. There must exist a certain constraint to keep the third position of codons G+C-rich in vertebrate genes. It should be noted that the pattern characteristic for vertebrates does not depend on the genes used; when codon usages of approximately 10 or more genes with varying functions were summed up for each vertebrate, they usually resulted in the converged pattern listed in table 5, regardless of differences in the genes compiled.

The finding that among taxonomically related organisms the codon-usage patterns summed for individual organisms resemble each other but that they differ between distant organisms is consistent with the "genome hypothesis" of Grantham et al. (1980, 1981). It should be mentioned, however, that the characteristic pattern of each multicellular organism usually becomes evident only after summing up genes with varying functions. In other words, when codon-choice patterns of individual genes (even those of one organism) are compared with each other, they are often very different. This seems to distinguish the codon-choice patterns of multicellular organisms from those of unicellular organisms. Examples of such diverse codon-choice patterns among human genes are listed in table 5. The third position of the codons of the zeta-globin gene and the skeletal-actin gene are primarily G or C, and their G+C percentages at the position are 95% and 89%, respectively. In contrast, the third position in the blood Christmas factor–IX gene and the alpha-fetoprotein gene is A+T-rich (35% and 38% G+C, respectively). Therefore, it is difficult to say that the codon-choice patterns of these two types of genes belong to a common dialect. In an effort to extensively study this diversity in G+C percentage at the third position, most of the sequenced vertebrate genes (~200 genes) have been examined, and the G+C-rich genes (>80% G+C) and the A+T-rich genes (<50% G+C) are listed in table 6. The highest G+C percentage was 98% (for the chicken histone H2A gene) and the lowest was 35% (for human blood Christmas factor–IX gene). Genes coding for the abundant cellular "housekeeping" proteins (e.g., actin, histone, and tubulin) usually have a high G+C percentage at this position, and, very interestingly, cellular oncogenes also have a high G+C percentage. By contrast, genes for blood-plasma proteins usually, but not always, have a low G+C percentage. Although the exact relationship of G+C

Table 6
G+C Percentage at the Third Position of Codons in Vertebrate Genes

G+C Percentage	Gene
G+C-Rich at the Third Position	
98	Histone 2A, 2B (cHIS2A2B)
95	GlobinZ (hHBA1)
93	Arginine Vasopressin (bAVPNPII)
91	Histone 4 (cHIS43), TubulinA (cTUBA)
90	TubulinB (cTUBB)
89	ActinA (hACTAS), GlobinA (hHBA2C)
89	MetallothioneinI (r, mMETI)
88	GlobinRH (cHBRH), Histone 5 (cHISH5)
88	Somatostatin (fSOMIAF)
86	POMC (bPOMC), Histone 1 (cHIS11A1)
85	ApolipoproteinA1 (hAPOA1)
84	Metallothionein2 (hMETII)
83	Histone 2A (cHISH2AF), HCGB (hCGBA)
82	GH (bGH), GlobinA, B (cHBAA, BM)
81	Rennin (bCHYMOAB), OncT (hONCT24)
80	Insulin1 (hINS155)
A+T Rich at the Third Position	
35	Christmas Factor (hFIX)
38	A-fetoprotein (hAFPA)
39	Albumin (hALB)
39	FibrinogenG (hFBRG)
40	AmylaseA1 (mAMY1M)
40	HormonPTH (hPTH2)
41	AmylaseA2 (mAMY2M)
42	Hormone PTH (bPTH2)
42	FibrinogenA (hFBRA)
42	SBP (rPSBPC)
42	HPRT (mHPRT)
44	InterferonG (hIFNG)
47	Ovalbumin (cOVAL)
47	KininogenL (bKIN1LMW)
49	OvalbuminFY (cY)

NOTE.—POMC = proopiomelanocortin; HCGB = chorionic gonadotropin; GH = growth hormone; SBP = steroid binding protein; HPRT = hypoxanthine phosphoribosyltransferase. GenBank locus names are listed in parentheses with abbreviated names of source organisms; b = bovine (BOV in GenBank); c = chicken (CHK); f = fish (FSH); h = human (HUM); m = mouse (MUS); and r = rat (RAT).

percentage and gene function is not known, we have recently found that genes with a high G+C percentage at the third position are usually surrounded by large sequences of high G+C content (Ikemura et al. 1983). The bar diagrams of figure 5 show such examples. In this figure, reported DNA sequences were divided into segments of 20 bases and the G+C percentages of the segments were sequentially arranged; starting and ending points of transcription are indicated by the arrows (↑) and exons by the dashes (——) under each diagram. In the case of the human

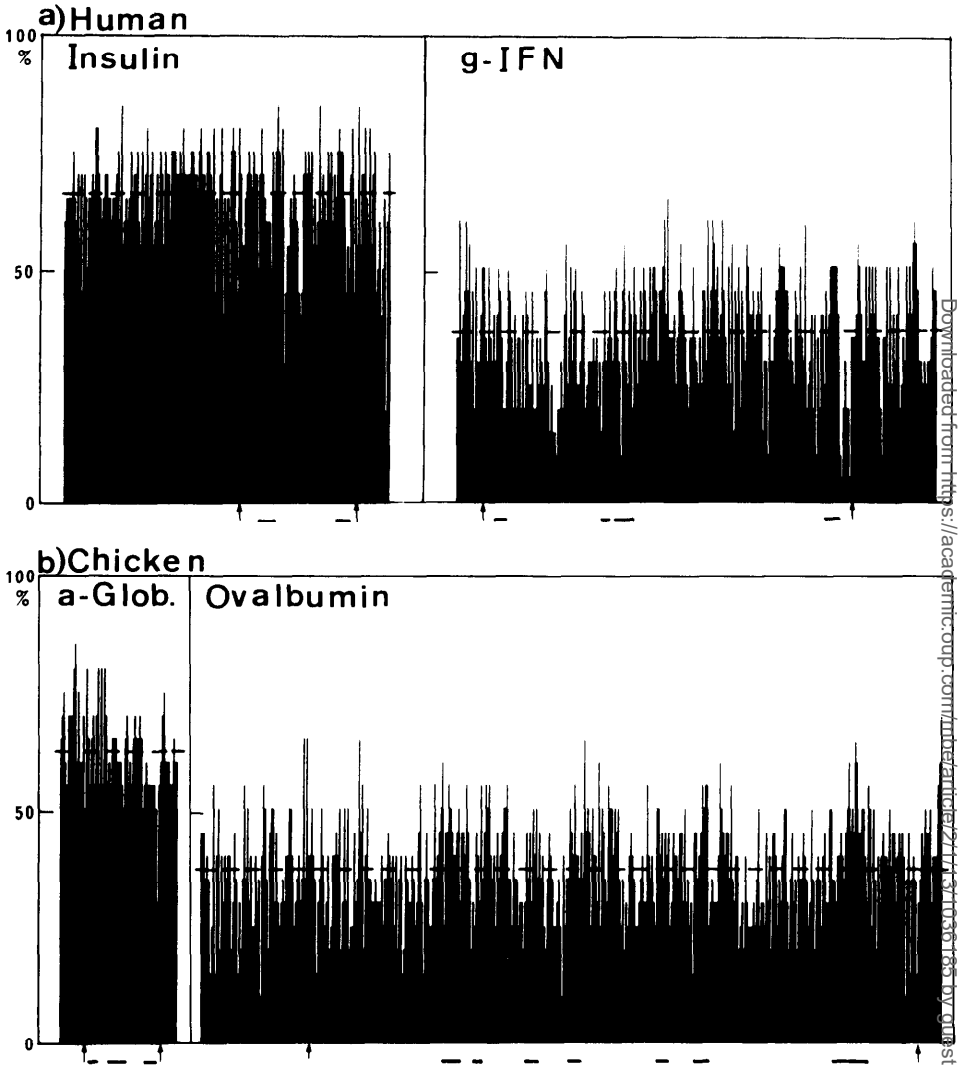

FIG. 5.—Distribution of G+C percentage in both the protein coding regions and the regions flanking them. The corresponding genome portion for each of the four vertebrate genes was divided into segments of 20 bases, and the G+C percentage in each segment is represented by the vertical bar. Segments are sequentially arranged from the 5′ side of the gene to the 3′ side. The G+C percentage in the total portion is shown by the horizontal broken line within the diagram. The initiation and termination sites of transcription are noted by the arrow (↑) under the diagram, and each exon is represented by a horizontal line (——) there. Therefore, the space between the exon lines corresponds to an intron. Represented are (*a*) human insulin (GenBank HUMINS1, 4,044 bp) and g-IFN (gamma interferon; HUMIFNG, 5,961 bp) genes and (*b*) chicken a-Glob. (alpha globin; CHKHBADA1, 1,468 bp) and Ovalbumin (CHKOVAL, 9,206 bp) genes.

insulin gene, which has a high G+C content (80%) at the third position, not only exons but also introns and the genome portions flanking the gene have a high G+C base composition. In contrast, the exons of the human gamma-interferon gene (g-IFN), whose G+C percentage at the third position is low (44%), are surrounded by A+T-rich flanking portions and by A+T-rich introns. This was also true for a pair

of chicken genes, the alpha-globin gene (82% G+C at the third position) and the ovalbumin gene (47% G+C) (figure 5b). Since DNA sequencing has mainly been focused on portions of structural genes or areas near the genes, there are only approximately 20 vertebrate genes around which a wide genome portion has been sequenced. Analyzing most of the available data, we again found a close correlation between the G+C percentage at the third position of codons and the G+C percentage of both the introns and the wide genome portion flanking the gene, at least at the upstream side (Aota, Ikemura, and Ozeki, unpublished data). Genes with a high G+C percentage at the third position (e.g., >75% G+C) are usually embedded in a wide genome portion of high G+C percentage (60%–70%), and genes with a low G+C percentage at the third position (i.e., <50%) are embedded within a wide portion with a low G+C percentage (usually <45%). It is interesting to note that large genome portions surrounding protein genes (even larger than the coding region—e.g., ~8 kb for ovalbumin) are kept either G+C-rich or A+T-rich, in accord with the G+C content of the third positions. This may be related to (1) some gross genetic information on gene expression lying in the large (e.g., several kb) DNA segment flanking the coding region and/or (2) the regional chromosome structure. Although the exact biological meaning of the variation of G+C-percentage distribution throughout the genome is unclear, this variation should be one causative factor in generating peculiar codon-choice patterns, resulting in either extremely G+C-rich or A+T-rich third positions and therefore also resulting in the observed diversity in G+C percentage at this position.

Distinction of Codon Choices between Unicellular and Multicellular Organisms

Concerning codon-choice patterns, I have emphasized the distinction between unicellular and multicellular organisms rather than the distinction between prokaryotes and eukaryotes. The reason will now be explained. Table 5 shows that by summing up codon usages for individual vertebrates, their codon choices converge to an essentially identical pattern. The factor responsible for this converging pattern might be related to the organisms' tRNA populations. The clearest example of the correlation between tRNA population and codon usages in multicellular organisms has been presented by Garel and his colleagues (Chevallier and Garel 1979; Garel 1982). In the posterior silk gland of *Bombyx mori,* which produces massive amounts of fibroin, the tRNA population (including the isoacceptor population) changes to conform to the codon frequencies of the fibroin gene at the end of the larval stage, and in the middle silk gland, the tRNA population does the same to conform to the codon usage in the sericin gene (Chevallier and Garel 1979). The authors therefore proposed a "functional adaptation of tRNA population to codon frequency." For unicellular organisms, I have emphasized the adjustment of codon choices to tRNA populations (Ikemura 1981a, 1981*b*). I think the difference lies in the distinction between unicellular and multicellular organisms. In the cell of a unicellular organism, a large number of genes are usually expressed. To keep their translation process efficient, these genes have an analogous codon-choice pattern (dialect) that fits the tRNA population. In contrast, in a highly differentiated cell of a multicellular organism, a restricted number of genes dominates the protein production of the cell. When some constraints on codon choice, arising from factors other than translation efficiency, exist for these genes, their codon choices are

presumably able to follow the constraints without loss of translation efficiency if the tRNA population in the cell adjusts to the codon choices of these principal genes, (as was found in the silk gland of *B. mori*). This freedom of codon choice in each highly differentiated cell—a freedom that is related to the ability to modulate its tRNA population without consideration of other genes—may bring about a diversity of codon-choice patterns for a multicellular organism. If this interpretation is correct, the converging pattern among vertebrates will be related at least in part to the average (or basal) tRNA populations of their cells, which presumably resemble each other among vertebrates. It should be noted, however, that at present it is unclear how general the phenomenon observed in the silk gland of *B. mori* is. Hastings and Emerson (1983) have proposed that the phenomenon in the silk gland may not be of general importance in vertebrates, showing that there are no tissue-specific differences in codon usages in muscle and liver genes of several vertebrates. It is conceivable that codon usages in the normal genes of multicellular organisms are less stringently constrained by tRNA content than are those in the genes of unicellular organisms, since the contribution, through translational efficiency change, of each synonymous mutation to the overall fitness change of a multicellular organism might be smaller than the corresponding contribution in a unicellular organism. The evident diversity of G+C percentage at the third position of codons among vertebrate genes (table 6) and the correlation of this G+C percentage with the G+C percentages of the flanking portions (fig. 5) may reflect this weaker constraint imposed by tRNA content. Further quantification of the tRNA population in various types of cells (or tissues) is clearly needed to learn how closely and generally codon usages and tRNA contents in multicellular organisms are related to each other.

In the present review, I have focused mainly on (1) the correlation found in unicellular organisms between codon usage and tRNA content and (2) the G+C percentage at the third position of codons in vertebrate genes, introducing both our work and that of other groups. Since various factors should have affected the codon-choice pattern during evolution, studies from many viewpoints are inevitably necessary for a fuller understanding of this process. In closing, I will refer to several recent reviews concerned with this topic. In an extensive compilation of codon-usage patterns in *E. coli* genes (83 genes), Gouy and Gautier (1982) have demonstrated the correlation between codon usage and gene expressivity and discussed the biological significance of the correlation. Grosjean and Fiers (1982) have discussed codon choices in light of codon-anticodon interaction and of translation efficiency and accuracy, and Hasegawa et al. (1979) have discussed it from the viewpoint of RNA secondary-structure requirements. Contextual constraint on codon choice (i.e., when the choice of codon is influenced by the neighboring codon) and dinucleotide preference rules have been proposed by Lipman and Wilbur (1983) and by Nussinov (1981), respectively.

Acknowledgments

The author is very grateful to Drs. H. Ozeki and M. Kimura for valuable discussion and encouragement; to Drs. T. Ooi, S. Aota, K. Kawasaki, and M. Ikemura for computer analysis; and to Dr. Dan Graur for critical reading of the manuscript and valuable comments. This work was supported by a grant-in-aid for scientific research from the Ministry of Education, Science and Culture of Japan (No. 58112008).

LITERATURE CITED

BENNETZEN, J. L., and B. D. HALL. 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

BONITZ, S. G., R. BERLANI, G. CORUZZI, M. LI, G. MACINO, F. G. NOBREGA, M. P. NOBREGA, B. E. THALENFELD, and A. TZAGOLOFF. 1980. Codon recognition rules in yeast mitochondria. Proc. Natl. Acad. Sci. USA **77**:3167–3170.

CHEVALLIER, A., and J. P. GAREL. 1979. Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in *Bombyx mori* by using two-dimensional polyacrylamide gel electrophoresis. Biochimie **61**:245–262.

CRICK, F. H. C. 1966. Codon-anticodon pairing: the wobble hypothesis. J. Mol. Biol. **19**:548–555.

GAREL, J. P. 1982. The silkworm, a model for molecular and cellular biologists. Trends Biochem. Sci. **7**:105–108.

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

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

GRANTHAM, R. 1980. Workings of the genetic code. Trends Biochem. Sci. **5**:327–331.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE, and R. MERCIER. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9**:r43–74.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER, and A. PAVE. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8**:r49–62.

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

GROSJEAN, H., D. SANKOFF, W. MINJOU, W. FIERS, and R. J. CEDERGREN. 1978. Bacteriophage MS2 RNA: a correlation between the stability of the codon:anticodon interaction and the choice of code words. J. Mol. Evol. **12**:113–119.

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

HASTINGS, K. E. M., and C. P. EMERSON, JR. 1983. Codon usage in muscle genes and liver genes. J. Mol. Evol. **19**:214–218.

HOPFIELD, J. J. 1974. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc. Natl. Acad. Sci. USA **71**:4135–4139.

IKEMURA, T. 1980. The frequency of codon usage in *E. coli* genes: correlation with abundance of cognate tRNA. Pp. 519–523 *in* S. OSAWA, H. OZEKI, H. UCHIDA, and T. YURA, eds. Genetics and evolution of RNA polymerase, tRNA and ribosomes. University of Tokyo Press, Tokyo; and Elsevier/North Holland, Amsterdam.

———. 1981*a*. 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.

———. 1981*b*. 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.

———. 1982. Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. **158**:573–597.

IKEMURA, T., S. AOTA, K. KAWASAKI, and H. OZEKI. 1983. Codon choice pattern of higher eukaryotes. Jpn. J. Genet. **58**:648.

IKEMURA, T., and H. OZEKI. 1977. Gross map location of *Escherichia coli* transfer RNA genes. J. Mol. Biol. **117**:419–446.

———. 1983. Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. Cold Spring Harbor Symp. Quant. Biol. **47**:1087–1097.

KANEHISA, M. I. 1982. Los Alamos sequence analysis package for nucleic acids and proteins. Nucleic Acids Res. **10**:183–196.

KIMURA, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624–626.

———. 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.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KING, J. L., and T. H. JUKES. 1969. Non-Darwinian evolution. Science **164**:788–798.

LIPMAN, D. J., and W. J. WILBUR. 1983. Contextual constraints on synonymous codon choice. J. Mol. Biol. **163**:363–376.

NICHOLS, B. P., M. BLUMENBERG, and C. YANOFSKY. 1981. Comparison of the nucleotide sequence of *trpA* and sequences immediately beyond the *trp* operon of *Klebsiella aerogenes, Salmonella typhimurium* and *E. coli*. Nucleic Acids Res. **9**:1743–1755.

NISHIMURA, S. 1978. Modified nucleosides and isoaccepting tRNA. Pp. 168–195 *in* S. ALTMAN, ed. Transfer RNA. MIT Press, Cambridge, Mass.

NUSSINOV, R. 1981. Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J. Mol. Biol. **149**:125–131.

PEDERSEN, S., P. L. BLOCH, S. REEH, and F. C. NEIDHARDT. 1978. Patterns of protein synthesis in *E. coli*. Cell **14**:179–190.

POST, L. E., and M. NOMURA. 1980. DNA sequences from the *str* operon of *Escherichia coli*. J. Biol. Chem. **255**:4660–4666.

POST, L. E., G. D. STRYCHARZ, M. NOMURA, H. LEWIS, and P. P. DENNIS. 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **76**:1697–1701.

THOMPSON, R. C., D. B. DIX, R. B. GERSON, and A. M. KARIM. 1981. A GTPase reaction accompanying the rejection of Leu-tRNA$_2$ by UUU-programmed ribosomes. J. Biol. Chem. **256**:81–86.

THOMPSON, R. C., and P. J. STONE. 1977. Proofreading of the codon-anticodon interaction on ribosomes. Proc. Natl. Acad. Sci. USA **74**:198–202.

WEISSENBACH, J., and G. DIRHEIMER. 1978. Pairing properties of the methylester of 5-carboxymethyl uridine in the wobble position of yeast tRNA. Biochim. Biophys. Acta **518**:530–534.

YANOFSKY, C., and M. VANCLEEMPUT. 1982. Nucleotide sequence of *trp*E of *Salmonella typhimurium* and its homology with the corresponding sequence of *Escherichia coli*. J. Mol. Biol. **155**:235–246.

MASATOSHI NEI, reviewing editor

Received April 11, 1984; revision received August 3, 1984.