Prokaryotic genetic code

S. Osawa, A. Muto, T. Ohama, Y. Andachi, R. Tanaka * and F. Yamao **

Laboratory of Molecular Genetics, Department of Biology, School of Science, Nagoya University, Chikusa-ku, Nagoya 464-01 (Japan)

Summary. The prokaryotic genetic code has been influenced by directional mutation pressure (GC/AT pressure) that has been exerted on the entire genome. This pressure affects the synonymous codon choice, the amino acid composition of proteins and tRNA anticodons. Unassigned codons would have been produced in bacteria with extremely high GC or AT genomes by deleting certain codons and the corresponding tRNAs. A high AT pressure together with genomic economization led to a change in assignment of the UGA codon, from stop to tryptophan, in *Mycoplasma*.
Key words. Genetic code; directional mutation pressure; GC-content; unassigned codon; tRNA; codon capture, codon usage; *Mycoplasma*; mitochondria; *Micrococcus*; methanogens.

Introduction

This article reviews the current status of the prokaryotic genetic code from the standpoint of comparative usage of codons and anticodons, on a phylogenetic basis. Emphasis has been laid on the role of directional mutation pressure on the code, and on a codon reassignment in *Mycoplasma*. Other important topics such as the use of UGA for selenocysteine, tRNA identity, suppressors, context effects, etc. are not included. For these, readers should refer to other articles in this volume or elsewhere.

GC-contents of bacterial genomes

A phylogenetic tree of representative groups of organisms constructed using 5S rRNA sequences suggests that eubacteria first separated from the metabacteria/eukaryotes branch[9]. Several major groups of eubacteria exist which seem to have diverged in different directions in the early stages of bacterial evolution. *Halobacterium, Thermoplasma, Sulfolobus*, the methanogens, etc. (metabacteria or archaebacteria) form a unique group among the prokaryotes. These bacteria and eukaryotes separated after the eubacterial emergence[13]. Among eubacteria, the mean GC-contents of genomic DNA varies from 25% to 75%. The gram-negative bacteria and the gram-positive bacteria diverged first. Among the gram-positive bacteria, those with a low genomic GC-content such as *Bacillus subtilis* (GC: 43%), and *Mycoplasma* spp. (– 25%) are phylogenetically close, whereas those with a high genomic GC-content, such as *Micrococcus luteus* (74%), and *Streptomyces griseus* (73%), comprise another phylogenetic group. These two groups separated long ago. The gram-negative bacteria with intermediate GC, such as *Escherichia coli* (50%), *Serratia marcescens* (58%), etc. belong to the common gram-negative branch (fig. 1)[9].

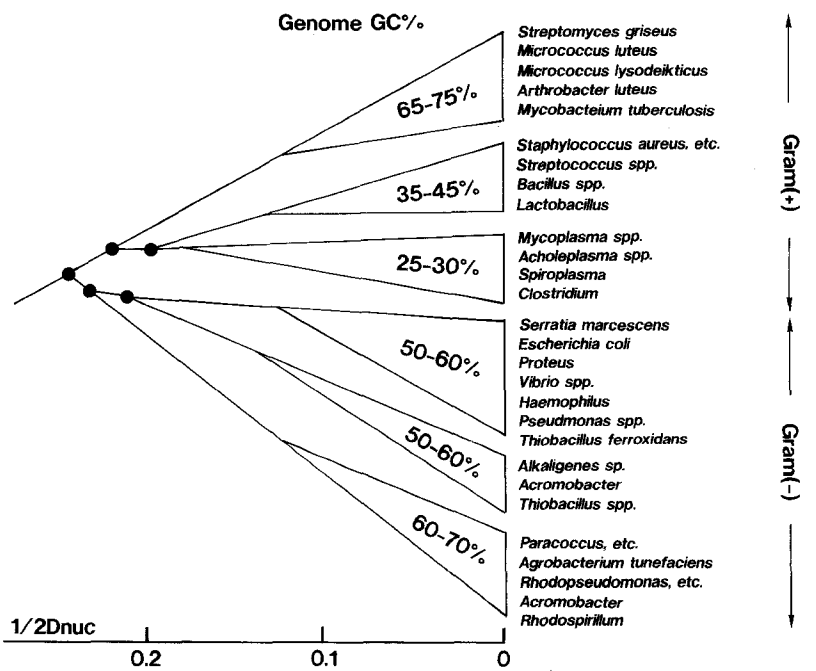

Figure 1. Phylogenetic tree of 5S rRNAs of representative eubacteria. *Dnuc*, relative evolutionary distance. Redrawn from Hori and Osawa[9].

Metabacteria may be classified into several groups; *Sulfolobus* and the methanogens are usually low in genomic GC (30–40%), while the genomes of *Halobacterium* spp. are rich in GC (63–67%)[9].

There is an argument that high genomic GC-content is a result of non-directional (random) mutations, followed by positive selection by temperature, because GC-pairs are more stable than AT-pairs in DNA[3,17]. However, this possibility is not very likely. Many GC-rich bacteria including *M. luteus* are not thermophilic, while a series of thermophiles has low GC-genomes. The genomic GC-contents are closely related to phylogeny and not to thermophilicity, at least in prokaryotes.

These variations in genomic GC-contents have been suggested to result from directional mutation pressure acting on the whole genome towards AT predominating over GC (AT-pressure), or towards GC predominating over AT (GC-pressure)[16,22,29]. It must be remembered, however, that GC/AT pressures are changeable upon, e.g., alteration of DNA duplication systems. This in turn explains why genomic GC-content varies in different phylogenetic lines.

The directional mutations are subject to selective constraints that generally eliminate functionally deleterious changes by negative selection. A certain fraction of non-deleterious mutations are then fixed in the population due to random genetic drift[18]. Thus, functionally less important parts of the genome evolve faster than more important parts. For a given species, GC/AT pressure changes the GC-content of various parts of the genome in the same direction, but to different extents, depending on their functional importance.

The GC-content of DNA is influenced by genes that are responsible for DNA replication. One example is the *mut-T* gene in *E. coli*[6]. It specially induces A:T → C:G transversions. Two genes in *E. coli*, *mut-M* and *mut-Y*, generate G:C → T:A transversions[5,23]. Evidently, there is enough variation in DNA-polymerase systems to provide for directional mutation pressure that will decide the base composition of DNA at various levels of GC-content.

Codon usage

The GC-content of the first, second and third codon positions of various bacterial species, when plotted against the corresponding genome GC-contents, reveals a linear positive relationship with genomic GC-contents, which has a different steepness of slope in the rank order of the third, first and second positions[27]. The most striking observation is that the GC-content of the third position varies linearly from about 10% in *Mycoplasma capricolum* to more than 90% in *Micrococcus luteus*. The fact that the correlation in the third position is strongest may be largely the result of GC/AT pressure, since the third positions and some of the first positions of codons are often synonymous between U and C, between A and G, or between U, C, A and G. This means that the percentage of G + C can range from 40% to 73% without any effect on amino acid content, simply by the selection of synonymous codons for amino acids ending in one case in U and A, and in the other case in C and G[14].

Table 1. Amino acid compositions of ribosomal proteins in the spc operon of *Mycoplasma capricolum*, *Escherichia coli*, and *Micrococcus luteus*

Amino acids	*M. capricolum* Mole %	*E. coli* Mole %	*M. luteus* Mole %
Phe	3.3	3.0	2.7
Leu	7.2	6.6	6.9
Ile	8.5	6.6	4.8
Met*	1.4	2.7	1.4
Val	8.5	9.6	10.6
Ser	6.1	4.2	4.4
Pro	3.1	3.3	4.1
Thr	6.5	4.9	6.9
Ala	7.3	10.7	9.5
Tyr	1.6	2.7	1.6
His	1.2	1.2	1.5
Gln	3.4	3.2	4.1
Asn	4.8	3.7	2.6
Lys	13.5	9.1	8.6
Asp	3.2	4.2	5.2
Glu	6.5	5.7	5.7
Cys	0.08	0.43	0.09
Trp	0.25	0.17	0.25
Arg	5.2	7.6	8.5
Gly	8.4	10.3	10.5
Total No. of codons	(1222)	(1154)	(1181)

Significantly increased and decreased amino acids in *M. luteus* as compared with *E. coli*/*M. capricolum* are underlined and hatched, respectively.

* Initiation codons are excluded.

Sueoka[28], before the genetic code was known, explored the relationship between GC-content of DNA and amino acid composition of total protein in a wide variety of bacterial species with GC in DNA ranging from 35% to 72%. Alanine and glycine were higher and isoleucine and tyrosine were lower in bacteria with a high GC-content than in their counterparts with low GC. This reflects the GC-content of the codons. This, in turn, means that directional mutation pressure also affects the amino acid composition of proteins. Its effect is exerted on amino acids in near-neutral sites in the protein molecule.

The average amino acid composition of eight homologous ribosomal proteins between *Mycoplasma capricolum* (genomic GC: 25%), *E. coli* (GC: 50%) and *Micrococcus luteus* (GC: 74%) (table 1) shows that the contents of phenylalanine, isoleucine, asparagine, and lysine decrease significantly with increasing genomic GC-content. These amino acids are assigned by codons which have A or U both at the first and second positions (UUY, AUY/A, AAY, and AAR, respectively; Y:U or C; R:A or G). On the other hand, the contents of amino acids assigned by codons having G or C both at the first and second positions, such as proline (CCN), alanine (GCN), arginine (CGN), and glycine (GGN), tend to increase with increasing genomic GC-content. Increases of valine and aspartic acid along with an increase in

Table 2. Anticodons in eubacteria and mitochondria

	L	E	M	mt		L	E	M	mt		L	E	M	mt		L	E	M	mt
Phe(UUU)	GAA	GAA	GAA	GAA	Ser(UCU)	GAA	GAA			Tyr(UAU)	GUA	GUA	GUA	GUA	Cys(UGU)	GCA	GCA	GCA	GCA
Phe(UUC)					Ser(UCC)			UGA	UGA	Tyr(UAC)					Cys(UGC)				
Leu(UUA)	–	UAA	UAA	UAA	Ser(UCA)	–*	UGA			–(UAA)					Trp(UGA)			UCA	UCA
Leu(UUG)	CAA	CAA	CAA		Ser(UCG)	CGA	CGA			–(UAG)					Trp(UGG)	CCA	CCA	CCA	
Leu(CUU)	GAG	GAG			Pro(CCU)	GGG	GGG			His(CAU)	GUG	GUG	GUG	GUG	Arg(CGU)				
Leu(CUC)			UAG	UAG	Pro(CCC)			UGG	UGG	His(CAC)					Arg(CGC)	ICG	ICG	ICG	U/ACG
Leu(CUA)	UAG	UAG			Pro(CCA)	–*	UGG			Gln(CAA)	–	UUG			Arg(CGA)				
Leu(CUG)	CAG	CAG			Pro(CAG)	CGG	CGG			Gln(CAG)	CUG	CUG	UUG	UUG	Arg(CGG)	CGG	CCG	–	
Ile(AUU)	GAU	GAU	GAU	GAU	Thr(ACU)	GGU	GGU	AGU		Asn(AAU)	GUU	GUU	GUU	GUU	Ser(AGU)	GCU	GCU	GCU	GCU
Ile(AUC)					Thr(ACC)			UGU	UGU	Asn(AUC)					Ser(AGC)				
Ile(AUA)	–	LAU	LAU	CAU	Thr(ACA)	UGU	UGA			Lys(AAA)	–*	UUU	UUU	UUU	Arg(AGA)	–	UCU	UCU	UCU
Met(AUG)	CAU	CAU	CAU		Thr(AUG)	CGU	CGU			Lys(AAG)	CUU		CUU	CUU	Arg(AGG)	CCU	CCU		
Val(GUU)	GAC	GAC			Ala(GCU)	GGC	GGC			Asp(GAU)	GUC	GUC	GUC	GUC	Gly(GGU)	GCC	GCC		
Val(GUC)			UAC	UAC	Ala(GCC)			UGC	UGC	Asp(GAC)					Gly(GGC)			UCC	UCC
Val(GUA)	–	UAC			Ala(GCA)	–*	UGC			Glu(GAA)	–*	UUC	UUC	UUC	Gly(GGA)	UCC	UCC		
Val(GUG)	CAC				Ala(GCG)	CGC				Glu(GAG)	CUC				Gly(GGG)	CCC	CCC		

Modifications of the anticodon first nucleoside are not indicated except for L, 2-lysyl-C (Ile) and I, inosine (Arg). L, *Micrococcus luteus* (unpublished); E, *Escherichia coli* (from ref. 19); M, *Mycoplasma capricolum* (from ref. 1); mt, fungal mitochondria (cited in ref. 1).
–, Neither tRNA nor the codons were found (probably unassigned codons); Not all anticodons were determined in L. –*, tRNA was not found, but may exist, because small amounts of the corresponding codons are used. In mt, both AUA and AUG = Met. In L and E, UGA = stop.

genomic GC-content are largely due to the conservative replacement of isoleucine (AUY/A) by valine (GUN), and of asparagine (AAY) by aspartic acid (GAY), respectively [24].

Anticodons of tRNA

Table 2 includes the anticodons in the eubacterial and mitochondrial codes as found or predicted in *Micrococcus luteus*, *E. coli*, *Mycoplasma capricolum*, and fungal mitochondria (Andachi et al.[1] and unpublished data). In table 3 are shown the known anticodons for *Halobacterium* spp. and methanogens [27]. The anticodon list is variable for different classes of bacteria. As discussed below, this variability is the result of the evolution of anticodons mainly as a result of GC/AT pressure, in response to the codon choice pattern.

The anticodon of a tRNA molecule pairs by hydrogen bonding with a codon in mRNA. Pairing between the second and third positions of anticodons with the second and first positions of codons follows the usual rules of A pairing with U and G with C. Crick [7] proposed that there is some play, or 'wobble' between the first anticodon base and the third codon base during pairing. As a result, anticodon GNN pairs with codon NNU as well as with codon NNC (N: U, C, A or G). Crick [7] also proposed that U pairs with A and G; I (inosine) with U, C and A; C only with G, and A, which is very rare in first anticodon positions, with U. Inosine occurs at the first anticodon position only in ICG for the arginine 'family box' in eubacteria. In addition, it is now evident that unmodified U pairs with all four bases, U, C, A and G in third positions of codons and that various modifications of U occur that restrict its pairing [27]. Codon-anticodon pairing rules in prokaryotes as deduced from table 1 are shown in figure 2.

A striking feature of the genetic code table is the presence of eight 'family boxes' in which four codons are assigned to the same amino acid. These eight family boxes are each potentially translatable by a single anticodon with the general formula UNN. This form of translation takes place in *Mycoplasma* [1] and some mitochondria [8]. In other bacteria, there is more than one anticodon per family box. One modification of U, **U, which is a 5-hydroxyuridine derivative (xo^{5}U) pairs with A, G and U [33]. Anticodons UNN and **UNN occur only in family boxes, and never in two-codon sets, because NNY and NNR in two-codon sets code for different amino acids. Reading of both NNY and NNR by a UNN or **UNN anticodon results in ambiguity of the NNY codons.

Twelve two-codon sets are not in family boxes. These sets end either in a pyrimidine (Y: U or C) or in a purine (R: A or G). Since NNY and NNR code for different amino acids, a distinction must be made between NNY codons pairing with anticodon GNN, and NNR codons pairing with anticodon UNN; in this case, the first anticodon

Table 3. Known anticodons in metabacteria

	H	M		H	M		H	M		H	M
Phe (UUU)	GAA	GAA	Ser (UCU)	GGA		Tyr (UAU)	GUA	GUA	Cys (UGU)	GCA	
Phe (UUC)			Ser (UCC)			Tyr (UAC)			Cys (UGC)		
Leu (UUA)	UAA		Ser (UCA)			Stop (UAA)			Stop (UGA)		
Leu (UUG)	CAA		Ser (UCG)	CGA		Stop (UAG)			Trp (UGG)	CCA	
Leu (CUU)	GAG		Pro (CCU)	GGG		His (CAU)	GUG	GUG	Arg (CGU)	GCG	
Leu (CUC)			Pro (CCC)			His (CAC)			Arg (CGC)		
Leu (CUA)	UAG	UAG	Pro (CCA)	UGG	UGG	Gln (CAA)		UUG	Arg (CGA)	UCG	
Leu (CUG)	NAG		Pro (CCG)	CGG		Gln (CAG)	CUG		Arg (CGG)	CCG	
Ile (AUU)	GAU		Thr (ACU)	GGU	GGU	Asn (AAU)	GUU	GUU	Ser (AGU)	GCU	
Ile (AUC)			Thr (ACC)			Asn (AAC)			Ser (AGC)		
Ile (AUA)	NAU	CAU	Thr (ACA)		UGU	Lys (AAA)	UUU	UUU	Arg (AGA)		UGU
Met (AUG)	CAU		Thr (ACG)	CGU		Lys (AAG)	CUU		Arg (AGG)		
Val (GUU)	GAC		Ala (GCU)	GGC		Asp (GAU)	GUC	GUC	Gly (GGU)	GCC	GCC
Val (GUC)			Ala (GCC)			Asp (GAC)			Gly (GGC)		
Val (GUA)		UAC	Ala (GCA)	UGC	UGC	Glu (GAA)	UUC	UUC	Gly (GGA)	UCC	
Val (GUG)	CAC		Ala (GCG)	CGC		Glu (GAG)	CUC		Gly (GGG)	CCC	

Halobacterium spp; M, methanogens: all anticodons are from *Methanococcus* spp. except GCC glycine which was reported only from *Methanobacterium*.

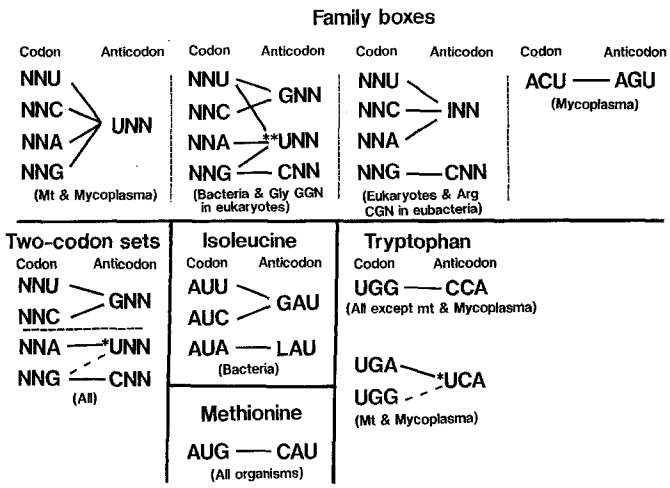

Figure 2. Codon-anticodon pairing rules. N: U, C, A or G; *U: 5-carboxymethylaminomethyl-uridine (cmnm^{5}U), 5-methyl-2-thiouridine derivative (xm^{5}s^{2}U), or 2'-O-methyluridine (Um); **U: 5-hydroxyuridine derivative (xo^{5}U); L: 2-lysyl cytidine (lysidine); I: inosine. -----: weak pairing. Drawn mainly from table 1 of Osawa and Jukes[27].

nucleotide, U, must be modified to prevent mispairing with NNY codons. In consequence, the U in all UNN anticodons in two-codon sets is modified. The seven pyrimidine-terminated sets are each translated by a single GNN anticodon, as discussed above. The five sets that end with purines are UUR, CAR, AAR, GAR and AGR, and they can each pair in translation with an anticodon *U, where *U is 5-carboxymethylamino methyluridine (cmnm^{5}U), 5-methoxycarbonylmethyluridine (mcm^{5}U) or a 2-thiolated uridine derivative[33]. These derivatives, especially the last one, pair strongly with A and only poorly with G. Because of the weak pairing ability of *U with G, NNR two-codon sets are sometimes translated by two anticodons; the G-terminated codon mainly by the anticodon CNN, and the A-terminated codon by the anticodon *UNN. The appearance of the anticodon *UNN in a family box is not harmful, as long as a GNN anticodon exists. In fact, tRNA with anticodon cmnm^{5}U is present in the glycine family box of *Bacillus subtilis*[12].

The isoleucine codon AUA in bacteria pairs with the anticodon *CAU, in which *C is 2-lysyl cytidine (lysidine: L)[21, 33].

The anticodon composition of *Halobacterium* spp. resembles the eubacterial code except for the absence of ICG and the presence of GCG and UCG. Since no CNN anticodons (except CAU) are included among the 17 known anticodons for methanogens, it seems likely that methanogens avoid the use of CNN anticodons[27].

As seen in table 4, a higher GC-content of bacteria is accompanied by an increase in GC in silent positions of codons, and since CNN anticodons translate only NNG codons, one would expect more CNN anticodons in bacteria with a higher GC-content. This is the case (table 2): *E. coli* (GC: 50%) has 12 CNN anticodons, and *Thermus thermophilus* (GC: 69%) has three more anticodons (CAC, CUU and CUC) that are not found in *E. coli*. This would also be the case in *M. luteus* (GC: 74%). *B. subtilis* (GC: 43%) has at least 6 CNN anticodons and

Table 4. Codon usage in *Micrococcus luteus*, *Escherichia coli* and *Mycoplasma capricolum*

	L	E	M		L	E	M		L	E	M		L	E	M
UUU(Phe)	< 1	19	39	UCU(Ser)	1	11	16	UAU(Tyr)	< 1	14	25	UGU(Cys)	< 1	4	6
UUC(Phe)	31	17	4	UCC(Ser)	30	10	< 1	UAC(Tyr)	25	14	4	UGC(Cys)	4	6	1
UUA(Leu)	0	10	64	UCA(Ser)	< 1	6	25	UAA(Stop)	–	–	–	UGA(Stop)	–	–	6*
UUG(Leu)	1	11	3	UCG(Ser)	13	7	1	UAG(Stop)	–	–	–	UGG(Trp)	4	11	1
CUU(Leu)	< 1	9	5	CCU(Pro)	2	6	10	CAU(His)	1	11	11	CGU(Arg)	10	28	8
CUC(Leu)	35	9	0	CCC(Pro)	17	3	< 1	CAC(His)	18	11	3	CGC(Arg)	50	21	1
CUA(Leu)	0	3	10	CCA(Pro)	< 1	7	18	CAA(Gln)	0	13	38	CGA(Arg)	1	3	< 1
CUG(Leu)	50	57	< 1	CCG(Pro)	27	25	< 1	CAG(Gln)	39	31	1	CGG(Arg)	15	4	0
AUU(Ile)	< 1	26	70	ACU(Thr)	1	11	30	AAU(Asn)	1	15	65	AGU(Ser)	< 1	6	17
AUC(Ile)	53	30	8	ACC(Thr)	43	24	1	AAC(Asn)	28	25	11	AGC(Ser)	4	15	3
AUA(Ile)	0	3	16	ACA(Thr)	< 1	6	22	AAA(Lys)	< 1	38	107	AGA(Arg)	0	1	28
AUG(Met)	22	26	22	ACG(Thr)	23	11	< 1	AAG(Lys)	51	12	10	AGG(Arg)	2	1	< 1
GUU(Val)	< 1	23	42	GCU(Ala)	2	19	33	GAU(Asp)	3	31	43	GGU(Gly)	9	31	26
GUC(Val)	44	14	1	GCC(Ala)	54	23	< 1	GAU(Asp)	51	23	4	GGC(Gly)	68	31	1
GUA(Val)	0	13	23	GCA(Ala)	3	21	21	GAA(Glu)	< 1	46	55	GGA(Gly)	2	5	29
GUG(Val)	52	25	2	GCG(Ala)	28	23	1	GAG(Glu)	73	19	4	GGG(Gly)	9	9	2

The numbers show the codon usage in the protein genes expressed in frequency per one thousand codons. < 1 represents the codon frequency less than 0.4. L, *Micrococcus luteus* (from ref. 24); E, *Escherichia coli* (from ref. 2); M, *Mycoplasma capricolum* (from refs 1, 25). *UGA is Trp codon in *M. capricolum*.

Mycoplasma capricolum (GC: 25%) has only five. Evidently AT pressure leads to a decrease in CNN anticodons, and CNN anticodons increase with GC pressure. This holds in metabacteria. Thirteen CNN anticodons have been found in *Halobacterium* spp. (GC: 63%) and only one in methanogens (GC: 30%), although two more (CAU and CCA) may be presumed to exist in methanogens[27].

Increases in GNN and CNN anticodons by GC-pressure would give more redundancy to the code and probably more fidelity in translation. Under extremely high GC-pressure, anticodon *U/**UNN or LAU (isoleucine) is reduced to trace amounts or becomes non-existent as a result of the decrease/disappearance of the corresponding NNA/NNU and AUA codons, respectively, as a result of GC-pressure. As GC-content decreases, GNN anticodons in family boxes also diminish in numbers, but they are necessarily retained in the other boxes for pairing with two-codon NNY sets. The eight family boxes in the *E. coli* code have 7 GNN anticodons, but these GNN anticodons are absent from *Mycoplasma capricolum* and a single UNN anticodon, the first nucleotide U, unmodified, pairing with four codons, is present in 6 of the 8 locations[1] (table 2).

The *M. capricolum* genome seems to have discarded the genes for many tRNAs, and those for many enzymes for tRNA nucleoside modifications, owing to the constraints for economizing its genome, which have resulted in the undermodified tRNAs[1]. Interestingly, *Acholeplasma laidlawii*, a member of the class Mollicutes, into which *Mycoplasma* is also classified, uses GNN anticodons together with **UNN anticodons in family boxes (unpublished). Since the genome size of *A. laidlawii* is twice that of *Mycoplasma*, the unique anticodon usage in *Mycoplasma* would have resulted from the genome economization that occurred in the *Mycoplasma* lineage after separation from the *Acholeplasma* lineage.

Mitochondrial tRNAs have several characteristic features resembling *M. capricolum* tRNAs. Most mitochondria (mt), except for green plants, use single UNN anticodons in all eight family boxes (table 2). Thus, the genomes of mt and mycoplasmas seem to have developed under similar evolutionary constraints, gene economization and AT-pressure, resulting in the similarities in their tRNAs.

Role of GC/AT pressure in species-specific codon usage

The codon usage or synonymous codon choice in prokaryotes reflects the GC-content of the genome as a whole as noted above. Another factor that affects the codon usage is the constraint imposed by tRNA.

It has been generally believed that the selective use of synonymous codons is influenced by tRNAs and is correlated with the degree of expression level of the genes. This phenomenon is seen in bacteria with a moderate genomic GC-content such as *E. coli*. Ikemura[10] has shown that the highly-expressed genes of *E. coli* reveal strongly biased codon usage of alternative synonymous codons, while the weakly expressed genes are less biased. For example, the NNU/C codons in two-codon sets are translated by a single anticodon GNN throughout bacteria and its pairing efficiency may be only somewhat better with the NNC codon than with NNU. Thus, in highly expressed genes of *E. coli* and *B. subtilis*, more NNC codons are chosen than NNU as a result of selection pressure by GNN anticodons to increase translation efficiency. On the other hand, there is no such demand for weakly expressed genes. Then the NNU to NNC ratio would not be affected by tRNAs and would be largely determined by mutation pressure. As shown in figure 3, there is a gradient of the NNC-contents of all the NNU/C-type two codon sets, decreasing from highly to weakly expressed genes, and reaching a 'bottom'-value of 40% on average in *E. coli* and 34% in *B. subtilis*. The 'bot-

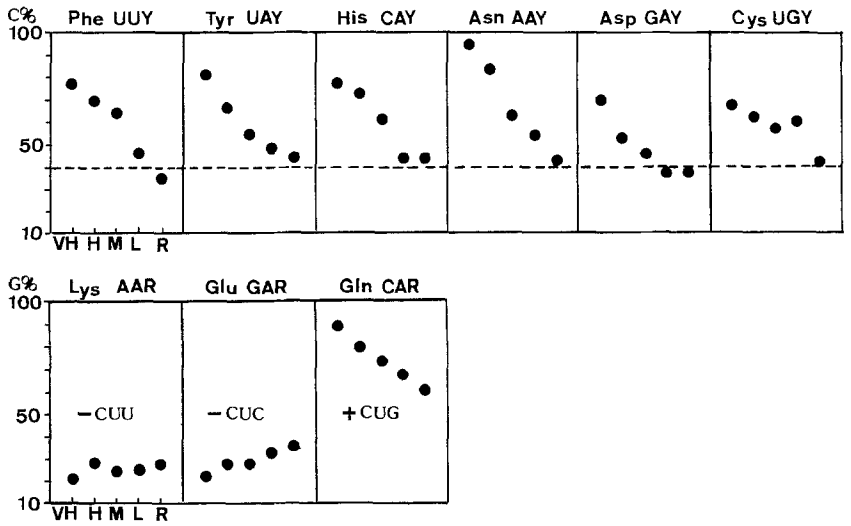

Figure 3. Codon usage in two-codon sets of various classes of genes in *Escherichia coli*. VH: very highly expressed genes; H: highly expressed genes; M: moderately expressed genes; L: genes expressed at a low level; R: regulatory genes expressed at a very low level. Dashed line: the average NNC-content of the class R genes of all the NNC/U-type two-codon sets shown in the figures. +/− CNN: presence/absence of CNN anticodon, from Ohama et al.[25].

tom'-value would thus represent the approximate degree of directional mutation pressure[25].

Generally speaking, the synonymous codon choice of the weakly expressed genes is much less affected by tRNA than that of the highly expressed genes, owing to a relaxation of selection by tRNA, but this does not mean that the silent positions are all free from such selection pressure. For example, in both *E. coli* and *B. subtilis*, contents of NNG codons in NNA/G-type two-codon sets (AAR lysine, and GAR glutamic acid) are less than 35% when translated by a single anticodon *UNN, regardless of expression levels[25] (fig. 3). The NNG content in *E. coli* is much less than that of GC-contents of whole genome (50%) or that of spacer regions (47%), suggesting the presence of 'negative' selection by *UNN anticodons to use fewer NNG codons than NNA throughout the genes. This would be because anticodon *UNN mainly pairs with codon NNA and very poorly with NNG, so that even in weakly expressed genes, NNG is a very 'bad' codon and cannot be used extensively. On the other hand, the presence of the CNN anticodon (e.g., in the CAR glutamine two-codon set), which translates exclusively codon NNG, strongly enhances the NNG usage with a gradient from the highly to the weakly expressed genes (fig. 3). Note that the NNG content in the weakly expressed genes is higher (60–70%) than the genomic GC-content, suggesting that 'positive' selection by CNN anticodons to use more NNG codons has been exerted even in this class of genes. On the contrary, in *M. luteus*, NNC/G codons are used almost exclusively and no difference in codon usage is shown between the highly and weakly expressed genes especially in all two-codon sets and isoleucine codons (fig. 4), suggesting that these codon usages are almost solely determined by GC mutation pressure, and not by selection pressure by tRNA.

The facts described above do not mean that all the codon usages of *M. luteus* are solely determined by mutation pressure without any selection pressure by tRNAs. In some family boxes, there are some differences in codon usage between the highly and weakly expressed genes.

In *M. luteus*, GUG is used as initiation codon in 10 out of 18 protein genes examined, whereas AUG much predominates over GUG in *E. coli*, or is exclusively used in *M. capricolum*. Most of the termination codons in *M. capricolum* and *E. coli* are UAA, whereas *M. luteus* uses 15 UGA, 2 UAA and 1 UAG[24]. Massive use of GUG initiation codons instead of AUG, and UGA termination codons instead of UAA in *M. luteus*, has most probably resulted from a strong preference for G usage over A, in accordance with the tendency to accumulate G/C at the silent positions.

We have seen that in *M. luteus*, choice of synonymous codons, especially in two-codon sets, is mainly determined by directional mutation pressure, and not by tRNA, because almost no constraints seem to exist between highly and weakly expressed genes. We have found that the extent of codon usage well parallels the amount of the corresponding tRNA. It is unlikely that GC-pressure first affects the *amount of tRNA* through which the codon usage is determined, since directional mutation pressure is exerted on individual nucleotide sites of DNA. It then follows that an increase/appearance or a decrease/disappearance of certain tRNA species is, in principle, an adaptive phenomenon affected by codon usage that has been primarily determined by directional mutation pressure. The tRNA population so produced would then modulate the codon usage by positive or negative selection. It is thus possible that high directional mutation pressure, in its extreme form, completely removes certain codons by converting them to other syn-

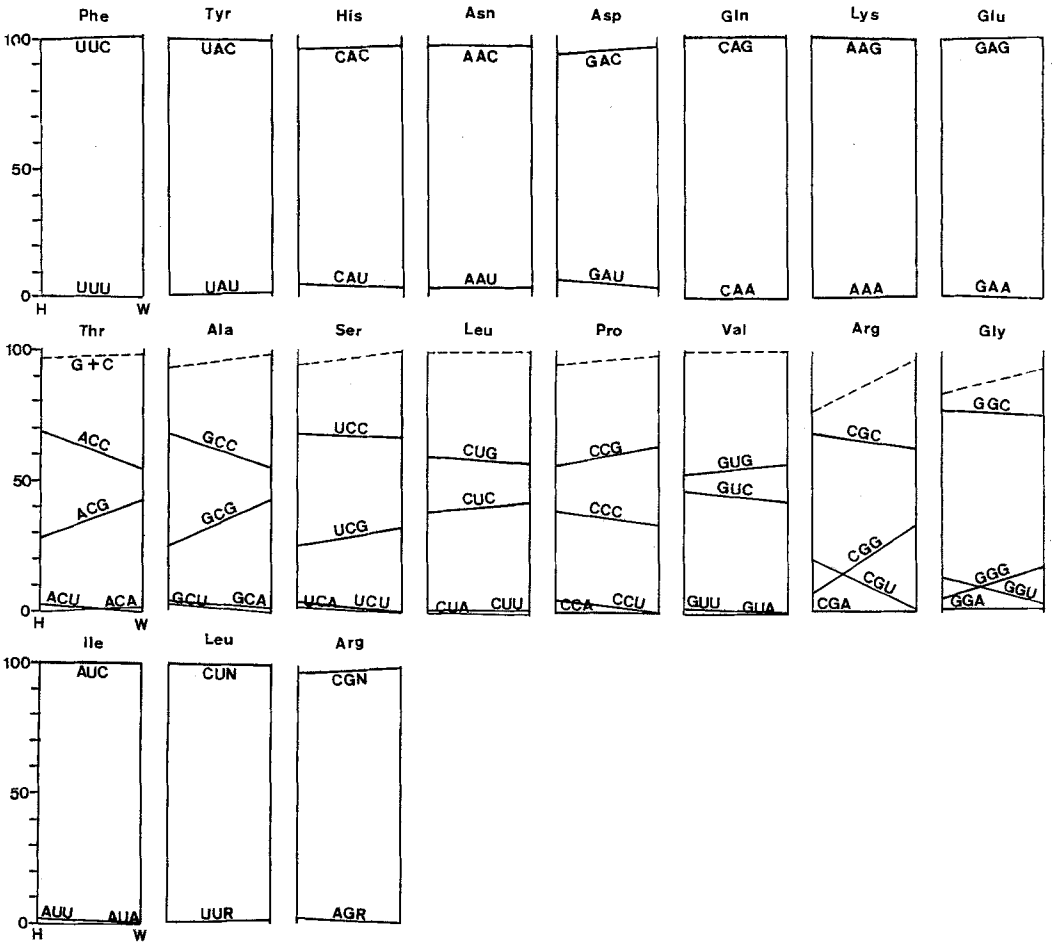

Figure 4. Synonymous codon ratios (%) in the highly and weakly expressed genes in *Micrococcus luteus*. H: highly expressed genes (ribosomal and its related protein genes). W: weakly expressed genes. -----: GC-contents of codon third positions of *M. luteus*. Cysteine (UGY), leucine (UUR), serine (AGY), and arginine (AGR) are not included in the figures of two-codon sets because of their small rate of occurrence, from Ohama et al.[25].

onymous codons, with concomitant disappearance of the corresponding tRNA; if a codon is withdrawn from use entirely, the corresponding tRNA gene will no longer be maintained by selective forces and may be lost from the genome, so that the codon will become unassigned.

NNA/G-type codons, except AUA (isoleucine), are translated primarily by a single anticodon *UNN (*U: includes both *U and **U from here on), with the occasional appearance in GC-rich bacteria of CNN anticodons pairing exclusively with NNG codons. Six codons, AUA (isoleucine), AGA (arginine), UUA (leucine), GUA (valine), CUA (leucine), and CAA (glutamine), all ending in A, are not found at all among 5516 codons in the *M. luteus* genes examined[24,25] (table 4). The tRNAs with anticodon *UNN responsible for reading these codons have not been detected either (unpublished data), suggesting that along with almost complete conversion of NNA codons to NNC or NNG by GC-pressure, *UNN anticodons are reduced to a trace amount or become non-existent. A few NNA codons that appear at this stage by mutation will be immediately selected against, because few *UNN anticodons exist to translate them. This is probably the reason for the almost complete absence of NNA codons in *M. luteus* genes. The absence of the AUA (isoleucine) codon may be explained by decrease or deletion of the corresponding anticodon *CAU (*C: lysidine (L)) along with silent conversion of the codon AUA to AUC by GC-pressure. It is then possible that of the undetectable NNA codons in *M. luteus* at least some may be unassigned codons.

By contrast, in eubacteria NNU and NNC codons are translated by a single anticodon GNN by wobble as noted above. Thus, a small amount of NNU codons that appeared as a result of mutations may be read by GNN anticodons, so that these codons are unable to become unassigned. In fact, a few NNU codons are used throughout in *M. luteus*.

A striking effect of AT-pressure can be seen in codon usage in *M. capricolum* genes[1,24] (table 4). More than 90% of the codon third positions are occupied by U or A, suggesting that many NNC and NNG codons have been converted to the synonymous NNU and NNA codons by AT-pressure. Remarkably, tRNA Arg (CCG), which is obligatory to read the CGG codon in eubacteria,

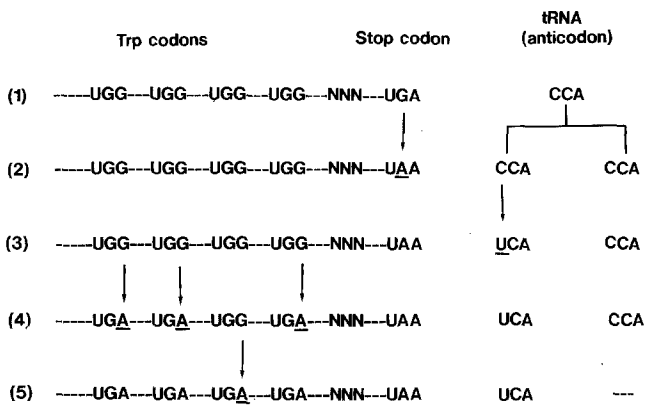

Figure 5. Reassignment of UGA codon from stop to tryptophan in the *Mycoplasma* lineage. Stages 1, 4 and 5 may be represented by *Acholeplasma laidlawii*, *Mycoplasma capricolum* and *Mycoplasma genitalium/pneumoniae*, respectively. For further explanations, see text.

is absent (table 2), and no CGG codon has been found in a total of 6814 codons examined in *M. capricolum* genes [1] (table 4). Perhaps strong AT-pressure has converted all CGG codons to the synonymous codons CGU, CGA or AGR by silent mutations. As a result, anticodon CCG has become unnecessary and has disappeared. Thus, CGG would have become an unassigned codon.

Three stop codons, UAA, UAG and UGA, do not exist in the coding frames in mRNA, except for very limited cases such as UGA for the selenocysteine codon. Thus, the stop codons would be able to become unassigned more easily than amino acid codons, by translation termination functions being restricted to one or two of the stop codons.

UGA is a tryptophan codon in Mycoplasma

Mycoplasma capricolum uses UGA, a universal stop codon, as a tryptophan codon [31]. In several other bacteria related to *M. capricolum* that have been investigated (three other *Mycoplasma* species [11] and *Spiroplasma citri* [4]), the UGA codon is also for tryptophan. *Acholeplasma laidlawii*, a member of the class Mollicutes, into which *Mycoplasma* and *Spiroplasma* are also classified, does not use UGA as a tryptophan codon [30]. It may be concluded that the code change, UGA for tryptophan, occurred in the *Mycoplasma/Spiroplasma* lineage after separation of these bacteria from *Acholeplasma* from their common ancestor.

It has been proposed that high AT-pressure in the *Mycoplasma* lineage resulted in the appearance of codon UGA for tryptophan [15, 31] (fig. 5). Many UGA and some UGG codons appear in the reading frames of the genes analyzed from these bacteria. Among them, a good number of UGA codons appear at the positions of tryptophan in the corresponding *E. coli* proteins. A simple conversion of the tryptophan codon UGG to UGA by AT-pressure could not happen, because UGA was an unusable codon, resulting from lack of anticodon UCA in the ancestor from which the *Mycoplasma* line branched. In fact, only one species of tRNA Trp (CCA) exists in *Acholeplasma* [30]. The conversion could happen only when tRNA Trp (UCA) appeared. Both tRNAs, tRNA Trp (UCA) which can translate both UGA and UGG, and tRNA Trp (CCA) have been found in this bacterium. The appearance of the UGA tryptophan codon would have taken place by the following steps without deleterious or lethal changes. AT-pressure has led to the replacement of all UGA stop codons by UAA. In AT-rich bacteria, the use of UGA as stop codons is rare, and most stop codons are UAA even in *E. coli* (GC: 50%). Therefore, complete conversion of UGA to UAA would not be so difficult in the *Mycoplasma* lineage. It is possible that along with this change, a releasing factor, which interacts both with UAA and UGA (RF-2 in *E. coli*), could have been deleted leaving RF-1 for UAA and UAG, or have become specific to UAA, so that UGA would have become an unassigned codon. Lee et al. [20] reported the presence of a release factor of the bacterial RF-1 type responding to UAA and UAG, and the absence of RF-2 for UGA, in rat mitochondria, where UGA is a codon for tryptophan as in *Mycoplasma*. It is thus possible that *Mycoplasma* would also have only RF-1, having discarded RF-2 during evolution, so that UGA is recognized only by tryptophan tRNA with anticodon UCA, and not by RF-2.

In *Mycoplasma*, following the changes described above, tryptophan tRNA, with anticodon CCA, duplicated, and one of the duplicates, under AT-pressure, mutated to UCA in its anticodon. The new tryptophan anticodon could pair with both UGA and UGG, so that UGA could be produced by AT-pressure causing mutation from UGG to UGA. The tandem arrangement of the genes for tRNA Trp (UCA) and tRNA Trp (CCA) on the chromosome of *M. capricolum* strongly supports that this was the case [15] (fig. 6). Since anticodon UCA and CCA both pair with the tryptophan codon UGG, this was a neutral change, because it was preceded by the disappearance of UGA stop codons, which had mutated to UAA. Anticodon CCA is no longer needed in *Mycoplasma*, and, although this is still present in *M. capricolum*, it is apparently disappearing [32]. The gene for tRNA Trp (CCA) has disappeared in some of the *Mycoplasma* species, such as *M. pneumoniae* and *M. genitalium* [11] (fig. 6). In this way, UGA would have been 'captured' (reassigned) by tryptophan in *Mycoplasma*.

We call this type of codon reassignment 'stop codon capture' because the former stop codon (in this case UGA) has been captured by an amino acid (tryptophan). A new tRNA [tRNA Trp (UCA)] is not a 'suppressor', since UGA is a regular tryptophan codon and is never used as a stop codon in this bacterium.

We have seen that there would be a number of unassigned codons in bacteria with extremely high GC/AT ratios. These unassigned codons may be reassigned by later reappearance of the codon and tRNA with a different

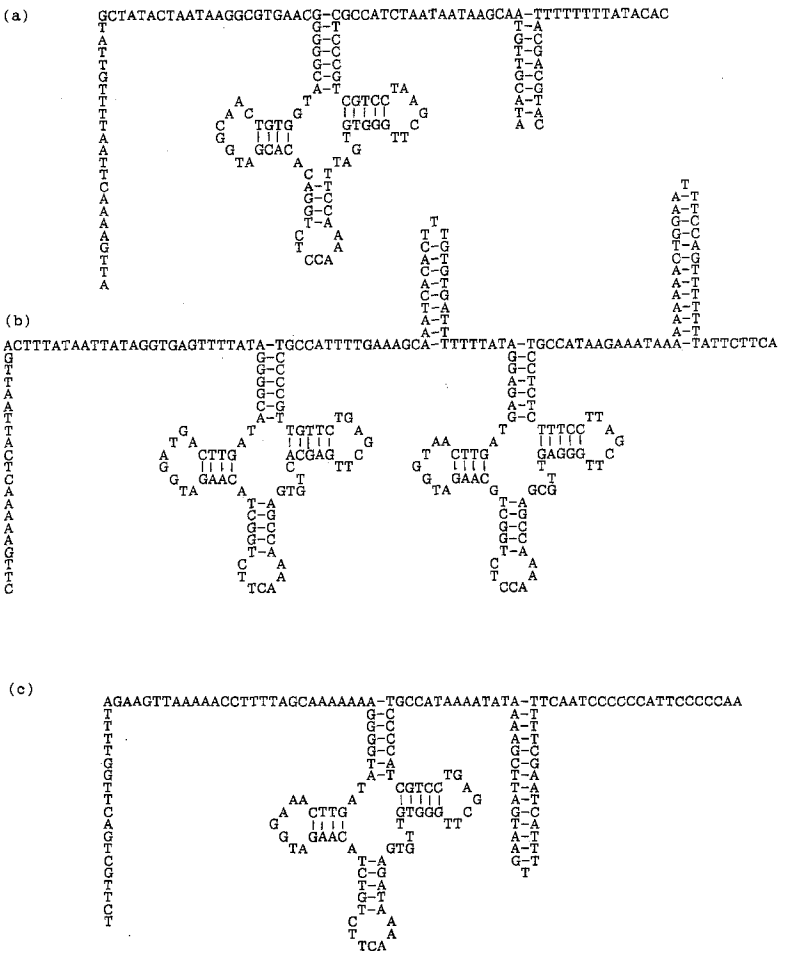

Figure 6. Tryptophan tRNA genes from Mollicutes. a) *Acholeplasma laidlawii*, taken from Tanaka et al.[30]; b) *Mycoplasma capricolum*, from Yamao et al.[31]; c) *Mycoplasma genitalium*, from Inamine et al.[11].

assignment (codon capture). A series of code changes known in mitochondria and eukaryotes may be explained by this codon capture hypothesis[27].

Conclusion

Most of the present-day organisms investigated use the 'universal' genetic code for protein synthesis. A series of changes in the genetic code, which deviates slightly from the universal code, has been recently discovered in most mitochondria and some of the organisms including *Mycoplasma* spp. These findings indicate that the universal code was not fixed, or 'frozen', in the ancestor from which all extant organisms are descended, but rather the code is still evolving in various lineages starting from the universal code.

In the present review, we have examined the codon and anticodon usages in prokaryotes. Codon and anticodon usages are influenced by directional mutation pressure. A strong AT-pressure apparently led to changes in the code in *Mycoplasma*. This change is interpreted as being caused by nondisruptive events such as the disappearance of a codon (in this case the UGA stop codon) and its reappearance with a different assignment (tryptophan).

The possible occurrence of unassigned codons would imply that the codon table for some life-forms would have less than 64 codons, and new changes will be discovered in various organisms. CGG is a good possibility for such a reassignment.

Acknowledgments. We thank Dr T. H. Jukes for valuable comments and Drs A. Böck and J. Auer for providing us anticodon composition of *Methanococcus*. This work was supported by a grant from the Ministry of Education, Science and Culture of Japan.

* On leave from Aburahi Laboratories, Shionogi & Ltd., Kouga, Shiga 520-34, Japan.

** Present address: National Institute of Genetics, Mishima, Shizuoka 411, Japan.

1 Andachi, Y., Yamao, F., Muto, A., and Osawa, S., Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *Mycoplasma capricolum*: Resemblance to mitochondria. J. molec. Biol. *209* (1989) 37–54.

2 Aota, S., Gojobori, T., Ishibashi, F., Maruyama, T., and Ikemura, T., Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. *16*, suppl. (1988) r315–r402.

3 Bernardi, G., and Bernardi, G., Compositional constraints and genome evolution. J. molec. Evol. *24* (1986) 1–11.

4 Bové, J. M., Carle, P., Garnier, M., Laigret, F., Renaudin, J., and Saillard, C., Molecular and cellular biology of spiroplasmas, in: The Mycoplasmas, vol. 5, pp. 243–364. Eds R. F. Whitcomb and J. G. Tully. Publishers, New York 1989.
5 Cabrera, M., Nghiem, Y., and Miller, J. H., *Mut*M, a second mutator locus in *Escherichia coli* that generates G.C → T.A transversions. J. Bact. *170* (1988) 5405–5407.
6 Cox, E. C., and Yanofsky, C., Altered base ratios in the DNA of an *Escherichia coli* mutator strain. Proc. natl Acad. Sci. USA *58* (1967) 1895–1902.
7 Crick, F. H. C., Codon-anticodon pairing: the wobble hypothesis. J. molec. Biol. *19* (1966) 548–555.
8 Heckman, J. E., Sarnoff, J., Alzner-DeWeerd, B., Yin, S., and Raj Bhandary, U. L., Novel features in the genetic code and codon reading patterns in *Neurospora crassa* mitochondria based on sequences of six mitochondrial tRNAs. Proc. natl Acad. Sci. USA *77* (1980) 3159–3163.
9 Hori, H., and Osawa, S., Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Molec. Biol. Evol. *4* (1987) 445–472.
10 Ikemura, T., Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. molec. Biol. *146* (1981) 1–21.
11 Inamine, J. M., Ho, K., Loechel, S., and Hu, P., Evidence that UGA is read as tryptophan rather than stop by *Mycoplasma pneumoniae*, *Mycoplasma genitalium* and *Mycoplasma gallisepticum*. J. Bact. *172* (1990) 504–506.
12 Ishikura, H., Murao, K., and Yamada, Y., EMBO-FEBS meeting, Strasbourg, July 1980.
13 Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T., Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. natl Acad. Sci. USA *86* (1989) 9355–9359.
14 Jukes, T. H., The genetic code, II. Am. Sci. *53* (1965) 477–487.
15 Jukes, T. H., A change in the genetic code in *Mycoplasma capricolum*. J. molec. Evol. *22* (1985) 361–362.
16 Jukes, T. H., and Bhushan, V., Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J. molec. Evol. *24* (1986) 39–44.
17 Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yasuhara, T., Tanaka, T., and Oshima, T., High guanine plus cytosine content in the third letter of codons of an extreme thermophile. DNA sequence of the isopropylmalate dehydrogenase of *Thermus thermophilus*. J. biol. Chem. *259* (1984) 2956–2960.
18 Kimura, M., The Neutral Theory of Molecular Evolution. Cambridge University Press, 1983.
19 Komine, Y., Adachi, T., Inokuchi, H., and Ozeki, H., Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12. J. molec. Biol. *212* (1990) 579–598.
20 Lee, C. C., Timms, K. M., Trotman, C. N. A., and Tate, W. P., Isolation of a rat mitochondrial release factor. Accommondation of the changed genetic code for termination. J. biol. Chem. *262* (1987) 3548–3552.
21 Muramatsu, T., Yokoyama, S., Horie, N., Matsuda, A., Ueda, T., Yamaizumi, Z., Kuchino, Y., Nishimura, S., and Miyazawa, T., A novel lysine-substituted nucleoside in the first position of the anticodon of minor isoleucine tRNA from *Escherichia coli*. J. biol. Chem. *263* (1988) 9261–9267.
22 Muto, A., and Osawa, S., The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. natl Acad. Sci. USA *84* (1987) 166–169.
23 Nghiem, Y., Cabrera, M., Cupples, C. G., and Miller, J. H., The mutY gene: A mutator locus in *Escherichia coli* that generates G:C → T:A transversions. Proc. natl Acad. Sci. USA *85* (1988) 2709–2713.
24 Ohama, T., Muto, A., and Osawa, S., Spectinomycin operon of *Micrococcus luteus*: Evolutionary implications of organization and novel codon usage. J. molec. Evol. *29* (1989) 381–395.
25 Ohama, T., Muto, A., and Osawa, S., Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content. Nucleic Acids Res. *18* (1990) 1565–1569.
26 Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F., and Osawa, S., The ribosomal protein gene cluster of *Mycoplasma capricolum*. Molec. gen. Genet. *210* (1987) 314–322.
27 Osawa, S., and Jukes, T. H., Evolution of the genetic code as affected by anticodon content. Trends Genet. *4* (1988) 191–198.
28 Sueoka, N., Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc. natl Acad. Sci. USA *47* (1961) 1141–1149.
29 Sueoka, N., On the genetic basis of variation and heterogeneity of DNA base composition. Proc. natl Acad. Sci. USA *48* (1962) 166–169.
30 Tanaka, R., Muto, A., and Osawa, S., Nucleotide sequence of tryptophan tRNA gene in *Acholeplasma laidlawii*. Nucleic Acids Res. *17* (1989) 5842.
31 Yamao, F., Muto, A., Kawauchi, Y., Iwami, M., Iwagami, S., Azumi, Y., and Osawa, S., UGA is read as tryptophan in *Mycoplasma capricolum*. Proc. natl Acad. Sci. USA *82* (1985) 2306–2309.
32 Yamao, F., Iwagami, S., Azumi, Y., Muto, A., Osawa, S., Fujita, N., and Ishihama, A., Evolutionary dynamics of tryptophan tRNAs in *Mycoplasma capricolum*. Molec. gen. Genet. *212* (1988) 364–369.
33 Yokoyama, S., Watanabe, T., Murao, K., Ishikura, H., Yamaizumi, Z., Nishimura, S., and Miyazawa, T., Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. Proc. natl Acad. Sci. USA *82* (1985) 4905–4909.

0014-4754/90/11-12/1097-10$1.50 + 0.20/0

Eucaryotic codes

F. Caron

Laboratoire de Génétique Moléculaire, Ecole Normale Supérieure, 46 rue d'Ulm, F-75230 Paris Cedex 05 (France)

Summary. This article is a review of the rules used by eucaryotic cells to translate a nuclear messenger RNA into a polypeptide chain. The recent observation that these rules are not identical in two species of a same phylum indicates that they have changed during the course of evolution. Possible scenarios for such changes are presented.
Key words. Genetic code; eucaryotic cell; evolution; code ambiguity; code universality; convergence.

Introduction

The genetic program of a cell is entirely contained in its DNA. Execution of this program involves many different steps, but each of them proceeds invariably in two phases which, for a eucaryotic cell, take place first in its nucleus and thereafter in the cytoplasm: in the nucleus, a copy of a part of the DNA is made and is afterwards modified by capping, polyadenylation, splicing and editing; in the cytoplasm, this modified copy (the messenger RNA) is