The genetic code – Thawing the 'frozen accident'

Dieter Söll[1,2,*] and Uttam L RajBhandary[3]

[1]*Department of Molecular Biophysics and Biochemistry,*
[2]*Department of Chemistry, Yale University, New Haven, CT 06520-8114, USA*
[3]*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

**Corresponding author (Fax, 1 (203) 432-6202; Email, soll@trna.chem.yale.edu)*

The sixties were an exciting period in molecular biology. The early studies of DNA and RNA synthesis had led to the discovery of DNA polymerase, DNA-dependent RNA polymerase and polynucleotide phosphorylase. Studies on a cell-free amino acid incorporating system, on transfer RNA, and the ribosome gave rise to much experimental effort to determine the mechanism of protein synthesis. All these developments focused high attention on deciphering the entire alphabet of the genetic code. Earlier work by Crick and Brenner (Crick *et al* 1961) had established that the code was a non-overlapping triplet code, and the stage was set for assigning the 64 codons to the 20 canonical amino acids. This was the subject of intense research in the laboratories of Marshall Nirenberg (e.g. Nirenberg and Leder 1964), Severo Ochoa (e.g. Lengyel *et al* 1961) and Har Gobind Khorana (e.g. Nishimura *et al* 1965). A critical advance was the discovery by Matthaei and Nirenberg that *Escherichia coli* extracts programmed with polyU directed the formation of poly phenylalanine; thus UUU was a codon for Phe (Nirenberg and Matthaei 1961). This then provided the translation approach for codon determination based on the analysis of the polypeptide products derived from defined synthetic mRNAs. In one such study the repeating dipeptide $(\text{Ser-Leu})_n$ was generated from a poly UC mRNA made by DNA-dependent RNA polymerase transcription of the DNA-like poly TC:AG sequence that was originally derived from short chemically synthesized oligonucleotides (Nishimura *et al* 1964, 1965). A second advance was the ribosomal binding technique (Nirenberg and Leder 1964) that proved to be crucial for the complete analysis of the code; it required only trinucleoside diphosphates (not long RNA molecules), a good preparation of ribosomes, and radioactive aminoacyl-tRNA. Fortunately, a good deal of knowledge existed on tRNA preparation and fractionation, and on the basic enzymatic properties of the aminoacyl-tRNA synthetases (RajBhandary and Köhrer 2006; Giegé 2006). Furthermore, the 20 canonical amino acids were available in radioactive form. The crucial 'triplets' (trinucleoside diphosphates) could be synthesized chemically (Lohrmann *et al* 1966) or enzymatically (Leder *et al* 1965) utilizing polynucleotide phosphorylase. Through such studies the genetic code assignments were established in 1965 (Nirenberg *et al* 1965, Söll *et al* 1965).

When it became clear that tRNAs possessed many modified nucleosides, present even in the anticodon, the question of tRNA decoding of the genetic code was addressed. For instance yeast tRNA^{Ala}, the first tRNA to be sequenced, was shown to have inosine as the first anticodon base (Holley *et al* 1965). This led Crick to propose the wobble hypothesis (Crick 1966) that suggested relaxed rules of pairing the third base of the codon with the first base of the anticodon; as a consequence a single tRNA species could recognize up to three codons (e.g. inosine would pair with U, C, and A in the third codon position). The confirmation came from triplet-binding experiments with fractionated tRNA species (Söll *et al* 1966) and protein synthesis experiments using purified tRNAs of known sequence (Söll and RajBhandary 1967). A further refinement to the genetic code came with the discovery of initiator tRNA, where it was shown that initiation of protein synthesis by initiator tRNA^{Met} (Marcker and Sanger 1964) uses not only AUG, but also GUG and UUG (Clark and Marcker 1966; Ghosh *et al* 1967).

At the time of its elucidation the genetic code was suggested to be universal in all organisms, and the result of a "frozen accident" unable to evolve further even if the current state were suboptimal (Crick 1968). Since alterations in the genetic code change the meaning of a codon, they cause unfaithful translation of the genetic message. While this was possibly acceptable at an early time in code development, at the stage of the current code "no new amino acid could be introduced without disrupting too many proteins" (Crick 1968). Thus it was assumed that the genetic code had no further evolvability. These and diverging early views of the genetic code and thoughts about the evolution of the translation system were more fully discussed at that time by Woese (1967).

How do we see the genetic code today – forty years after the familiar ‘alphabet’ was established? Is the code still viewed as a frozen accident or has a thaw set in? Of course, the “genetic code” (usually presented as the correlation of mRNA codons and amino acids) is the product of its interpretation by the translational machinery and it is only static as long as the components of this machinery do not change/evolve or are strictly conserved between organisms. The components foremost involved in this process are the tRNA molecules whose anticodon matches the codon of the mRNA by the rules of the wobble hypothesis (Crick 1966; Söll *et al* 1966), the aminoacyl-tRNA synthetases (aaRSs) which ensure correct acylation of each tRNA species with its cognate amino acid (Ibba and Söll 2000), and the peptide chain termination factors that govern the use of the termination codons (Kisselev *et al* 2003).

It is therefore not surprising that genetic and biochemical studies of the translation machinery over the past four decades and more recently genome analyses have challenged the concept of a non-evolving code. The results of these investigations illustrate that the genetic code is still evolving even though fixation of mutations that lead to codon reassignment is not favoured. A diverse assortment of alterations in the genetic code have now been documented with changes in 10 codons in nuclear codes and 16 codons in mitochondrial codes (reviewed in Osawa *et al* 1992; Knight *et al* 2001; Santos *et al* 2004; Miranda *et al* 2006). Many of these changes involve the termination (stop) codons UAA, UAG, and UGA.

In the altered nuclear codes UAA and UAG are used as glutamine codons in some green algae, ciliates (e.g. *Tetrahymena*), and *Diplomonads*. UGA is more versatile: it encodes cysteine in *Euplotes*; tryptophan in some ciliates, *Mycoplasma* species, *Spiroplasma citri* and *Bacillus*; and an unidentified amino acid in *Pseudomicrothorax dubius* and *Nyctotherus ovalis*. The leucine codon CUG also encodes serine in *Candida* species and in many *Ascomycetes.* The codons AGA (arginine) and AUA (isoleucine) in *Micrococcus* species (Kano *et al* 1993) and CGG (arginine) in *Mycoplasma capricolum* (Oba *et al* 1991) are either extremely rare or absent, and appear not to encode an amino acid. This provided the reason for Genoscope (Paris, France) to put *Micrococcus luteus* on their list of genomes to be sequenced.

The altered mitochondrial genetic codes are even more diverse. Mitochondria of many phyla with the exception of green plants encode tryptophan with UGA. UAA is used for tyrosine in some species of *Platyhelminth*, while UAG encodes alanine and leucine in the mitochondria of some plants and fungi. Metazoan and *Saccharomyces* mitochondria use AUA (isoleucine) as a methionine codon. In *Platyhelminth* and *Echinoderm* mitochondria AAA (normally lysine) encodes asparagine. The canonical arginine codons AGA and AGG specify stop in vertebrate mitochondria and serine in some animal mitochondria. In a few organisms AGA encodes glycine, and serine in some others. In mitochondria of the green alga *Scenedesmus obliquus* UCA (normally serine) is a stop codon.

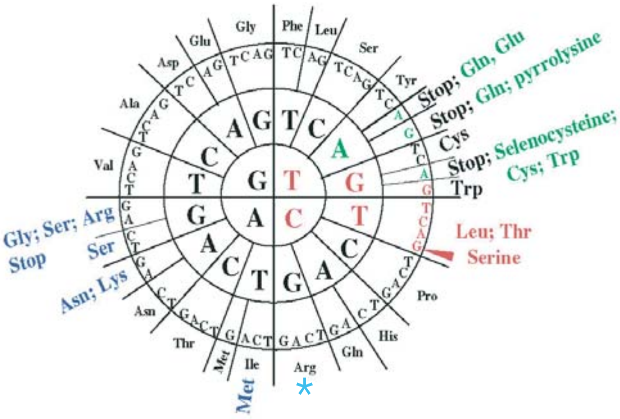

Figure 1. Summary of genetic code changes. Unchanged codons are in black. Blue: only mitochondria; green: bacteria, eukaryotes or mitochondria; red: yeast cytoplasm and mitochondria. The blue asterisk indicates the CGN codons which may be unassigned in yeast mitochondria. (Adapted from Miranda *et al* 2006).

All these changes are summarized in figure 1. It is obvious that most codons are unchanged from the ‘universal’ code. The figure makes it clear that a smaller number of codons gives rise to these evolutionary code changes in many different organisms. Apart from the stop codons UAA, UAG and UGA, the single codons AUA, AAA, AGA, AGG, CUG as well as the two four-codon boxes CUN and CGN are mutable in their meaning.

Changing the genetic code: How did these variations arise? Three explanations have been attempted. The *codon capture hypothesis* (Osawa and Jukes 1989) postulates that during genome evolution a codon together with the tRNA species carrying the corresponding anticodon disappears. Later the codon may again emerge together with the appropriate anticodon. However, the tRNA carrying this anticodon may be specific for an amino acid different from that associated with the tRNA that disappeared. Such a codon reassignment is called “codon capture”, and this phenomenon is explained as the result of biased G+C or A+T pressure. The *ambiguous intermediate hypothesis* (Schultz and Yarus 1994) claims that tRNA species with dual identity (capable of being charged by two different aaRSs) bring about genetic code changes through a gradual codon identity change. The process is problematic, since decoding ambiguity reduces an organism’s fitness. However, the theory has good support, as the CUG codon in certain *Candida* species is ambiguous and the decoding $tRNA_{CAG}$ is charged both by seryl-tRNA synthetase (SerRS) and leucyl-tRNA synthetase (LeuRS)

(Suzuki *et al* 1997). As a matter of fact, ambiguous decoding is an inherent feature of mRNA translation; the standard decoding error rate is 10^{-4}. The *genome streamlining hypothesis* (Andersson and Kurland 1995) suggests that code change, at least in mitochondria, is driven by the pressure to shrink the translation apparatus during genome reduction. The hypothesis associates codon reassignments in animal mitochondria with the reduced tRNA populations (see below) in these organelles.

More recent discoveries have revealed code changes derived from the evolution of amino acids leading to two additional co-translationally inserted amino acids. Selenocysteine, the 21st amino acid found in selected organisms from all three domains of life, is encoded by UGA (Böck *et al* 2005). Pyrrolysine, the 22nd amino acid found mainly in the *Methanosarcinaceae*, is inserted in response to a UAG codon (Srinivasan *et al* 2002). Both cases underscore that stop codon reassignments are preferred in the expansion of the genetic code.

*tRNAs – Adaptors of the genetic code***:** What tRNA complement does it take to read the genetic code? The wobble hypothesis predicted (Crick 1966) and its experimental verification showed (Söll *et al* 1966) that a single tRNA can recognize up to three codons; thus translation of the 61 sense codons requires a minimum of only 32 tRNA species. The minimal number of tRNAs decoding the mitochondrial genome is even smaller (24 tRNAs in *Neurospora*) as an unmodified U in the first position of the anticodon can read all four codons in a box (Breitenberger and Rajbhandary 1985). Moreover, mammalian mitochondria have only 22 tRNA species, as there is only one tRNA^{Met} (serving both as initiator and elongator tRNA) and AUA is decoded as methionine (with no necessity for a second tRNA^{Ile}). However the codon:anticodon pairing rules have been expanded by the recognition that anticodon conformation is shaped by the anticodon loop/stem structure (Yarus 1982) which is critically dependent on nucleotide modification (reviewed in Agris 2004). Most modifications are found in the first anticodon position (position 34) and in the base following the anticodon (position 37); they exert their effect either on codon recognition, or on the efficiency of codon reading, or on both. Some of these observations are summarized in table 1. Most excitingly, the molecular details of some of these expanded wobble interactions have now been visualized in crystal structures of codon-anticodon pairs in the decoding site of the 30S ribosomal subunit (Murphy *et al* 2004).

Accurate and efficient translation also depends on codon usage. Different organisms, influenced by their genome's G+C content, make use of the redundancy of the genetic code to use a biased selection of codons for certain amino acids during protein synthesis. Codon usage in prokaryotes and eukaryotes is well correlated with the amount of the particular tRNA isoacceptor in the cell (Ikemura 1985; Kanaya *et al* 2001) and appears to be related to tRNA gene dosage. This might be important for high level gene expression, especially for heterologous genes that possess a different codon frequency. The knowledge, that efficient decoding of the rare *Escherichia coli* codons AGA (arginine), AUA (isoleucine), CUA (leucine), and CCC (proline) may be problematic, has led to the construction of *E. coli* strains suitable for high-level heterologous protein expression (e.g. BL21-CodonPlus containing extra copies of the *E. coli argU, ileY, leuW*, and *proL* tRNA genes; Stratagene, La Jolla, CA, USA).

Since the genetic code is 'interpreted' by the translational machinery, it is not surprising that most of the non-standard codes arise from tRNA changes generated by post-transcriptional modifications (table 1) rather than by base substitutions in tRNA anticodons. Thus, tRNA modification is an essential component of an organism's code evolution process. This might explain why a large amount of a genome's coding capacity (estimated to be about 5% in *E. coli*) is dedicated to RNA modification. The evolutionary process is based on some plasticity in an organism's coding response and on the fact that a somewhat destabilized

Table 1. Some tRNA modifications and their effect on codon recognition.

1st anticodon base	3rd codon position recognized	Species or organelle
U	U, C, A or G	Mitochondria, chloroplasts, Mycoplasma
xo^5U	U, A or G	Bacteria
xm^5U	A or G	Bacteria, eukaryotes, mitochondria
$\text{xm}^5\text{s}^2\text{U}$	A only	Bacteria, eukaryotes
k^2C (L)	A only	Bacteria
m^7G	U, C, A or G	Echinoderm, squid mitochondria
f^5C	A or G	Mitochondria of *Drosophila*, nematode, bovine

N, any nucleotide; xo^5U, 5-hydroxymethyluridine derivative; xm^5U, 5-methyluridine derivative; $\text{xm}^5\text{s}^2\text{U}$, 2-thio-5-hydroxymethyluridine derivative; k^2C (L), lysidine; m^7G, 7-methylguanosine; f^5C, 5-formyl cytidine. For structures see Motorin and Grosjean (1998).

proteome can be tolerated by the cell, even though such conditions are also detrimental to the cell.

Based on our current understanding of the mechanism and accuracy of translation, a thought-provoking theory of collective evolution very recently suggested that non-Darwinian mechanisms present in early communal life (e.g. horizontal gene transfer of translational components) may account for the optimality and universality of the genetic code (Vetsigian *et al* 2006). In light of these developments it is clear that a better understanding of the translation apparatus derived from current and future biochemical, genetic and genomic approaches will shed more light on the evolution of the genetic code.

Acknowledgments

This article is dedicated to Har Gobind Khorana, our postdoctoral advisor (1962-1967) at the Enzyme Institute of the University of Wisconsin. In his laboratory we learned the excitement and demands of conducting science at the cutting edge. During that time we fell in love with the subject of our scientific careers, *transfer RNA*. Its pursuit allowed us to stay in lifelong contact with some of our friends from the Khorana laboratory: Susumu Nishimura, Eiko Ohtsuka, Hikoya Hayatsu, Günter Weimann and Simon Chang.

We thank Michael Ibba and Sotiria Palioura for critical reading of the manuscript. Current work in our laboratory on tRNA is supported by grants GM22854 (to DS) and GM67741 (to ULR) from the National Institute of General Medical Sciences.

References

Agris P F 2004 Decoding the genome: A modified view; *Nucleic Acids Res.* **32** 223–238

Andersson S G and Kurland C G 1995 Genomic evolution drives the evolution of the translation system; *Biochem. Cell Biol.* **73** 775–787

Böck A, Thanbichler M, Rother M and Resch A 2005 Selenocysteine; in *Aminoacyl-tRNA Synthetases* (eds) M Ibba, C S Francklyn and S Cusack (Georgetown, TX: Landes Bioscience) pp 320–327

Breitenberger C A and Rajbhandary U L 1985 Some highlights of mitochondrial research based on analyses of *Neurospora crassa* mitochondrial DNA; *Trends Biochem. Sci.* **10** 478–483

Clark B F and Marcker K A 1966 The role of N-formyl-methionyl-sRNA in protein biosynthesis; *J. Mol. Biol.* **17** 394–406

Crick F H, Barnett L, Brenner S and Watts-Tobin R J 1961 General nature of the genetic code for proteins; *Nature* (*London*) **192** 1227–1232

Crick F H C 1966 Codon–anticodon pairing: The wobble hypothesis; *J. Mol. Biol.* **19** 548–555

Crick F H C 1968 The origin of the genetic code; *J. Mol. Biol.* **38** 367–379

Ghosh H P, Söll D and Khorana H G 1967 Studies on polynucleotides. LXVII. Initiation of protein synthesis in vitro as studied by using ribopolynucleotides with repeating nucleotide sequences as messengers; *J. Mol. Biol.* **25** 275–298

Giegé R 2006 The early history of tRNA recognition by aminoacyl-tRNA synthetases; *J. Biosci.* **31** 477–488

Holley R W, Apgar J, Everett G A, Madison J T, Marquisee M, Merrill S H, Penswick J R and Zamir A 1965 Structure of a ribonucleic acid; *Science* **147** 1462–1465

Ibba M and Söll D 2000 Aminoacyl-tRNA synthesis; *Annu. Rev. Biochem.* **69** 617–650

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

Kanaya S, Yamada Y, Kinouchi M, Kudo Y and Ikemura T 2001 Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis; *J. Mol. Evol.* **53** 290–298

Kano A, Ohama T, Abe R and Osawa S 1993 Unassigned or nonsense codons in *Micrococcus luteus*; *J. Mol. Biol.* **230** 51–56

Kisselev L, Ehrenberg M and Frolova L 2003 Termination of translation: Interplay of mRNA, rRNAs and release factors?; *EMBO J.* **22** 175–182

Knight R D, Freeland S J and Landweber L F 2001 Rewiring the keyboard: Evolvability of the genetic code; *Nat. Rev. Genet.* **2** 49–58

Leder P, Singer M F and Brimacombe R L 1965 Synthesis of trinucleoside diphosphates with polynucleotide phosphorylase; *Biochemistry* **4** 1561–1567

Lengyel P, Speyer J F and Ochoa S 1961 Synthetic polynucleotides and the amino acid code; *Proc. Natl. Acad. Sci. USA* **47** 1936–1942

Lohrmann R, Söll D, Hayatsu H, Ohtsuka E and Khorana H G 1966 Studies on polynucleotides. LI. Syntheses of the 64 possible ribotrinucleotides derived from the four major ribomononucleotides; *J. Am. Chem. Soc.* **88** 819–829

Marcker K and Sanger F 1964 N-formyl-methionyl-sRNA; *J. Mol. Biol.* **12** 835–840

Miranda I, Silva R and Santos M A 2006 Evolution of the genetic code in yeasts; *Yeast* **23** 203–213

Motorin Y and Grosjean H 1998 Chemical structures and classification of posttranscriptionally modified nucleosides in RNA; in *Modification and editing of RNA* (eds) H Grosjean and R Benne (Washington: American Society for Microbiology) pp 543–549

Murphy F V, 4th, Ramakrishnan V, Malkiewicz A and Agris P F 2004 The role of modifications in codon discrimination by $tRNA^{Lys}_{UUU}$; *Nat. Struct. Mol. Biol.* **11** 1186–1191

Nirenberg M and Leder P 1964 RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of sRNA to ribosomes; *Science* **145** 1399–1407

Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F and O'Neal C 1965 RNA codewords and protein synthesis, VII. On the general nature of the RNA code; *Proc. Natl. Acad. Sci. USA* **53** 1161–1168

Nirenberg M W and Matthaei J H 1961 The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides; *Proc. Natl. Acad. Sci. USA* **47** 1588–1602

Nishimura S, Jacob T M and Khorana H G 1964 Synthetic deoxyribopolynucleotides as templates for ribonucleic acid polymerase: The formation and characterization of a ribopolynucleotide with a repeating trinucleotide sequence; *Proc. Natl. Acad. Sci. USA* **52** 1494–1501

Nishimura S, Jones D S and Khorana H G 1965 Studies on polynucleotides. 48. The in vitro synthesis of a co-polypeptide containing two amino acids in alternating sequence dependent upon a DNA-like polymer containing two nucleotides in alternating sequence; *J. Mol. Biol.* **13** 302–324

Oba T, Andachi Y, Muto A and Osawa S 1991 CGG: An unassigned or nonsense codon in *Mycoplasma capricolum*; *Proc. Natl. Acad. Sci. USA* **88** 921–925

Osawa S and Jukes T H 1989 Codon reassignment (codon capture) in evolution; *J. Mol. Evol.* **28** 271–278

Osawa S, Jukes T H, Watanabe K and Muto A 1992 Recent evidence for evolution of the genetic code; *Microbiol. Rev.* **56** 229–264

RajBhandary U L and Köhrer C 2006 Early days of tRNA research: Discovery, function, purification and sequence analysis; *J. Biosci.* **31** 439–451

Santos M A, Moura G, Massey S E and Tuite M F 2004 Driving change: The evolution of alternative genetic codes; *Trends Genet.* **20** 95–102

Schultz D W and Yarus M 1994 Transfer RNA mutation and the malleability of the genetic code; *J. Mol. Biol.* **235** 1377–1380

Söll D, Jones D S, Ohtsuka E, Faulkner R D, Lohrmann R, Hayatsu H and Khorana H G 1966 Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique; *J. Mol. Biol.* **19** 556–573

Söll D, Ohtsuka E, Jones D S, Lohrmann R, Hayatsu H, Nishimura S and Khorana H G 1965 Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids; *Proc. Natl. Acad. Sci. USA* **54** 1378–1385

Söll D and RajBhandary U L 1967 Studies on polynucleotides. LXXVI. Specificity of transfer RNA for codon recognition as studied by amino acid incorporation; *J. Mol. Biol.* **29** 113–124

Srinivasan G, James C M and Krzycki J A 2002 Pyrrolysine encoded by UAG in archaea: Charging of a UAG-decoding specialized tRNA; *Science* **296** 1459–1462

Suzuki T, Ueda T and Watanabe K 1997 The 'polysemous' codon–a codon with multiple amino acid assignment caused by dual specificity of tRNA identity; *EMBO J.* **16** 1122–1134

Vetsigian K, Woese C and Goldenfeld N 2006 Collective evolution and the genetic code; *Proc. Natl. Acad. Sci. USA* **103** 10696–10701

Woese C R 1967 *The genetic code* (New York, NY: Harper and Row)

Yarus M 1982 Translational efficiency of transfer RNA's: Uses of an extended anticodon; *Science* **218** 646–652

*e*Publication: 4 October 2006