Inouye, Cell Dev Biol 2018, 7:1
DOI: 10.4172/2168-9296.1000192

Cell & Developmental Biology
ISSN: 2168-9296

Commentary Open Access

Mystery of the Genetic Code

Inouye M*

Department of Biochemistry and Cell Biology, Rutgers-Robert Wood Johnson Medical School, Center for Advanced Biotechnology and Medicine, Piscataway, New Jersey, USA

Abstract

There are a total of 64 genetic codons assigned to the 20 amino acids and termination codons. The fact that all living organisms share the same codon assignment for individual amino acids indicates that all living organisms on the earth originated from the same organism. Mysteriously, however, the number of codons for individual amino acids does not necessarily correlate to amino acid usages in currently living organisms. In this article, I will discuss this mystery of the genetic code.

Introduction

The genetic codon table was established in 1963 by Crick [1] as shown in Table 1. Evolutionarily, it remains a mystery how the genetic codons for the individual amino acids were established. In the very early stages of evolution or at the origin of life, there must have been more than the current 20 amino acids and possibly a mixture of different stereospecific forms, L and D. During the evolution, only the L forms of 20 specific amino acids were finally selected to be the components of proteins. In addition, for proteins production, the genetic codons for individual amino acids consist of three bases out of four different bases, thymidine, cytosine, adenine and guanine making a total of 64 different combinations for 20 different amino acids.

Analysis of the Codon Table

When the codon table shown in Table 1 is examined, one can immediately realize that the numbers of codons for individual amino acids are quite different from 1 to 6:

- Two amino acids are encoded by only one codon; tryptophan (Trp; UGG) and methionine (Met; AUG)
- Nine amino acids are encoded by two codons; phenylalanine (Phe; UUU and UUC), tyrosine (Tyr; UAU and UAC), cysteine (Cys; UGU and UGC), histidine (His; CAU and CAC), glutamine (Gln; CAA and CAG), asparagine (Asn; AAU and AAC), lysine (Lys, AAA and AAG), aspartic acid (Asp; GAU and GAC), glutamic acid (Glu; GAA and GAG)
- One amino acid is encoded by three codons; isoleucine (Ile; AUU, AUC and AUA)
- Five amino acids are encoded by four codons; proline (Pro; CCU, CCC, CCA and CCG), threonine (Thr; ACU, ACC, ACA and ACG), valine (Val; GUU, GUC, GUA and GUG), alanine (Ala; GCU, GCC, GCA and GCG), glycine (Gly; GGU, GGC, GGA and GGG)
- Three amino acids are encoded by six codons; serine (Ser; UCU, UCC, UCA, UCG, AGU and AGC), leucine (Leu; UUA, UUG, UCU, UCC, UCA and UCG), arginine (Arg; CGU, CGC, CGA, CGG, AGA and AGG)

	T	C	A	G
T	TTT Phe (F) TTC Phe (F) TTA Leu (L) TTG Leu (L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Try (W)
C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)
A	ATT Ile (I) ATC Ile (I) ATA Ile (I) **ATG** Met (M)	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)
G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)

Table 1: Genetic codon table [1].

The analysis of the codon table reveals the following three aspects

Amino acids of the similar properties (for example, acidic amino acids and hydrophobic amino acids) are arranged in the same row or column.

Codons for the amino acids are generally arranged so that a mutation by one base results in an amino acid of similar qualities; for example, Gly to Ala to Val to Ile to Leu (or the reverse direction) is GGU to GCU to GUU to AUU to CUU; Asn to Asp to Glu to Gln (or the reverse direction) is AAC to GAC to GAG to CAG; Phe to Tyr (or the reverse direction) is UUU to UAU, and Thr to Ser (or the reverse direction) is ACU/C/A/G to UCU/C/A/G or ACU/C to AGU/C. Interestingly, the Thr-to-Ser mutation can be achieved by two independent mutations; for example, ACU can be changed to either UCU or AGU.

There are a few amino acids, which do not follow rule #2 described above; Lys (AAA and AAG) and Arg (CGU, CGC, CGA, CGG, AGA and AGG) are basic amino.

Acids, which can be mutually exchangeable in proteins (not in all cases). However, if Arg is encoded by CGU, CGC, CGA or CGG, they cannot be converted to AAA or AAG by a single mutation. Similarly, Pro (CCU, CCC, CCA and CCG) and Gly (GGU, GGC, GGA and GGG) are mutually exchangeable (not in all cases), but at the codon level, two bases have to be changed for the Gly-to-Pro and Pro-to-Gly alterations.

***Corresponding author:** Inouye M, Department of Biochemistry and Cell Biology, Rutgers-Robert Wood Johnson Medical School, Center for Advanced Biotechnology and Medicine, Piscataway, New Jersey, USA, Tel: (848) 445-9813; E-mail: inouye@cabm.rutgers.edu

Received December 18, 2017; **Accepted** January 06, 2018; **Published** January 15, 2018

Citation: Inouye M (2018) Mystery of the Genetic Code. Cell Dev Biol 6: 192. doi:10.4172/2168-9296.1000192

Frequencies of Amino acid Usages in Proteins and the Numbers of Codons

One question is why the number of codons for individual amino acids differs from 1 to 6. If could be thought that these numbers are proportional to amino acid usage frequencies in proteins, since some amino acids such as Trp and Cys are generally minor amino acids in proteins, whereas amino acids such as Leu and Ser are highly used. In Table 2, the amino acid composition (%) in mammalian proteins is shown, cited from the paper by Gaur [2], showing that Leu, Ser and Asp are the top three amino acids in mammalian proteins, whereas Trp, Cys and Met are the least used amino acids. In Table 2, the number of codons for each amino acid is also shown so that the average load (La) for each codon for individual amino acids can be calculated:

La=amino acid composition (%)/the number of codons of that amino acid

For example, La for Leu=9.79/6=1.63 and La for Asp=4.93/2=2.47.

If the La number for amino acid A is three times higher than the La number for amino acid B, the codons for A are used on average three times more frequently in protein synthesis. Thus, one can find from Table 2 that the codons for Glu (E) are most frequently used (La=3.53), while the codons for Arg (R) are least used (La=1.04). However, one ought to be cautious interpreting these numbers; for example since Glu is encoded by two codons, it is likely that one codon is used more frequently than the other, indicating that the La number for Glu is higher than 3.54 for one codon, while the La number for the other codon is less than 3.54. In the same manner, as Arg is encoded by six different codons, some Arg codons may be used at the La value of higher than 3. The average La for all 20 amino acids is 1.75, so that those amino acids having La values higher than 1.75 are the rate limiting amino acids for protein synthesis. In particular, Glu (E) is the most rate limiting amino acid, having an La value of 3.53 for mammalian protein synthesis. Arg, Leu and Ser are encoded by six codons, having the La values of 1.04, 1.63 and 1.72, respectively, all of which are lower than the average La value (1.75). However, it is unknown at present which codons out of the six codons have the highest La values.

Amino Acid	Composition	La Value
Leu	9.7	1.63
Ile	4.38	1.46
Phe	3.66	1.83
Trp	1.22	1.22
Val	6.02	1.51
Met	2.03	2.03
Ala	6.94	1.74
Gly	6.86	1.72
Pro	6.51	1.63
Cys	2.14	1.07
Tyr	2.6	1.3
Thr	5.16	1.29
Glu	7.05	3.53
Ser	8.15	1.72
Gln	4.48	2.24
Asp	4.93	2.47
His	2.5	1.25
Asn	3.66	1.83
Lys	5.68	2.84
Arg	6.22	1.04

Table 2: The amino acid composition of non-membrane mammalian proteins [2] and the La value* of each amino acid.
*The average load (La) for each codon for individual amino acids can be calculated: La=amino acid composition (%)/the number of codons of that amino acid
For example, La for Leu is 9.70/6=1.63

Another important factor is that the rate of protein synthesis is regulated by the concentrations of charged tRNAs for individual amino acids. Even with a low La value for a particular amino acid, the rate of protein synthesis is not necessarily high, if the concentration of charged tRNA for that amino acid is low. Although no data are at present available on the concentrations of tRNAs for individual amino acids in cells, one can assume that tRNA concentrations are quite equal for all amino acids. In this case, the presence of the Glu codons slows down the rate of protein synthesis, acting as a regulator for protein production.

Evolution of Genetic Codons

Another question can be raised as to how the codon table was settled as shown in Table 1, as some amino acid codons violate the hypothesis that codons for similarly functional amino acids are arranged in the same row or line in the codon table. This ordering allows for a point mutation in a codon to not cause a major functional change in the amino acid encoded by the mutated triplet codon. For example, Pro and Gly share a similar function in proteins, as they could be mutually exchangeable in proteins. However, Pro is encoded by CCX (X=U, C, A or G), while Gly is encoded by GGX, so that the Pro-to-Gly mutation (or the reverse mutation) requires two base changes.

It has been speculated that evolutionarily, the first amino acids are Gly, Ser, Asp and Asn, which are encoded by GGX, UCX or AGY (Y=U or C), GAY and AAY, respectively [3]. Supposing that Gly is the most primitive amino acid as it has no side group, one can assume that poly-G could also be one of the most primitive poly amino acid chains. If this is the case, then how could Ser, Asp and Asn also be the most primitive amino acids considering their genetic codons. One has to consider the following mutational pathway; GGG (Gly) to GGC (Gly) to AGC (Ser) to AAC (Asn) to GAC (Asp) [the third base could be C or T] or GGG (Gly) to GGC (Gly) to GAC (Asp) to AAC (Asn) to AGC (Ser). As discussed above, Glu has the highest La value among all 20 amino acids in mammalian proteins, and the change from Gly to Glu could have occurred by one base change (the second G to A) GGG to GAG. As shown in Table 2, among the amino acids in mammalian non-membrane proteins, the content of Leu is the highest (9.79%), followed by Ser (8.15), Glu (7.05), Ala (6.94), Gly (6.86), Pro (6.51), Arg (6.22), Val (6.02), Lys (5.68) and Thr (5.16). Trp is the lowest with the composition of 1.22%. Interestingly, their La values are not in the same order as the amino acid composition; Leu (1.63), Ser (1.72), Glu (3.53), Ala (1.74), Gly (1.72), Pro (1.63), Arg (1.04), Val (1.52), Lys (2.34) and Thr (1.29). Thus, Glu has the highest La value followed by Lys, Ala, Gly=Ser, Pro=Leu, Val, Thr, Arg. Notably, among these 10 amino acids, Glu is the only amino acid encoded only by two codons. Three amino acids having 6 codons (Leu, Ser, Arg) are in this group, while all the others (Ala, Gly, Pro, Val and Thr) are encoded by four codons except for Lys, which is encoded by 2 codons.

It is interesting to note that all the amino acids in this group have lower La values than the average La value (1.74) except for Glu (3.53), suggesting that the Glu codons (GAA and GAG) seem to have a unique role in the regulation of the rate of protein synthesis in mammalian cells. The amino acids with genetic codons differing by only one base from the GGG codon of Gly are Glu (GAG), Ala (GCG), Val (GUG), Arg (AGG and CGG) and Trp (UGG). However, these amino acids do not match with the amino acids Gly, Ser, Asp and Asn which are speculated to be the first amino acids in evolution [2].

Another important consideration concerning the rate of protein synthesis is codon usage. Among the synonymous codons, there are substantial differences in their usages depending on the organism. Such nonrandom usage of synonymous codons has been shown to correlate with the relative quantities of individual tRNAs [4-6] so that those codons used the least are defined as the minor codons, while those used preferentially are termed the major codons. Thus, the genes encoding abundant proteins selectively use the major codons. In *E. coli*, it has been demonstrated that AGA and AGG are the minor codons for Arg among the six Arg codons and if these minor codons are used in the first 25 codons from the AUG initiation codon, the rate of protein synthesis is substantially reduced [7]. However, if the tRNA for these minor codons are overproduced or the minor codons are moved by more than 50 codons away from the initiation codon, the inhibitory effect of the minor codons is substantially reduced. The analysis of a total of 26,000 human coding sequences revealed that there is preferential usage of the minor codons for Ala, Pro, Ser and Thr in the initial 50 codons of the open reading frames [8]. The minor codons for these four amino acids share CG at the second and the third base such as GCG for Ala, CCG for Pro, UCG for Ser and ACG for Thr. Since by the use of these minor codons, the concentrations of tRNAs per codon becomes highest in comparison with the other synonymous codons, the translation efficiency is also considered to become highest at the minor codons [8]. This hypothesis was experimentally confirmed by comparing the expression of the luciferase gene encoded by minor codons with that encoded by major codons.

Other Factors Affecting the Efficiency of Protein Synthesis

In addition to codon usage, there are a number of other factors regulating the efficiency of protein synthesis such as the secondary structures in mRNAs, the GC content of mRNAs and the concentrations of tRNA synthases. As for the major mystery of the codon table, the choice of a group of codons for a particular amino acid could have been accidental at least for some amino acids and if so, there is no way to come to a logical explanation of the evolution of the current codon table. It is noted that some amino acids seem to specifically interact with their codons [9], however, since this stereo-chemical principle cannot be applied for most amino acids, the initial assignments of codons together with co-evolution and adaptation are considered to complete the current codon table as recently discussed by Yarus [10]. Since no organisms having proteins encoded by different codon assignment have be found on the earth, all the current living organisms are considered to be originated from the last organism using the current codon assignment, which was partially established based on chemical rationale and partially by random events.

References

1. Crick FH (1963) On the genetic code. Science 139: 461-464.
2. Gaur RK (2014) Amino acid frequency distribution among eukaryotic proteins. IIOAB J 5: 6-11.
3. Crick FH, Brenner S, Klug A, Pieczenik G (1976) A speculation on the origin of protein synthesis. Orig Life 7: 389-397.
4. Ikemura T (1981a) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146: 1-21.
5. Ikemura T (1981b) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151: 389-409.
6. Maruyama T, Gojobori T, Aota S, Ikemura T (1986) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res14: r151-197.
7. Chen G-FT, Inouye M (1994) Role of the AGA/AGG codons, the rarest codons in global gene expression in *Escherichia coli*. Gene Dev 8: 2641-2652.
8. Park JH, Kwon M, Yamaguchi Y, Firestein BL, Park JY, et al. (2017) Preferential use of minor codons in the translation initiation region of human genes. Hum Genet 136: 67-74.
9. Hobish MK, Wickramasinhhe NE, Ponnamperuma C (1995) Direct interaction between amino acids and nucleotides as a possible physicochemical basis for the origin of the genetic code. Adv Space Res 15: 365-382.
10. Yarus M (2017) The genetic code and RNA-amino acid affinities. Life (Basel) 7: pii: E13.