Scientific Research Publishing

Open Journal of Genetics, 2020, 10, 8-15
https://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

The Pattern of Occurrence of Cytosine in the Genetic Code Minimizes Deleterious Mutations and Favors Proper Function of the Translational Machinery

Bin Wang

Department of Chemistry, Marshall University, Huntington, WV, USA
Email: wangb@marshall.edu

How to cite this paper: Wang, B. (2020) The Pattern of Occurrence of Cytosine in the Genetic Code Minimizes Deleterious Mutations and Favors Proper Function of the Translational Machinery. *Open Journal of Genetics*, **10**, 8-15.
https://doi.org/10.4236/ojgen.2020.101002

Received: January 10, 2020
Accepted: February 29, 2020
Published: March 3, 2020

Abstract

The standard genetic code consists of 64 combinations of base triplets made from four different bases. The research aim of this study was to investigate the pattern of occurrence of cytosine in the genetic code. By exploring the base composition and sequence of all 64 codons, the author found some important features based on the instability of cytosine. Because cytosine undergoes spontaneous deamination that converts it into uracil, it is evolutionarily favorable to exclude cytosine from codons critical to the initiation and termination of translation. For amino acids that have one to three synonymous codons (also called synonyms), the frequency of occurrence of C in the first and second positions of their mRNA codons is significantly lower than the frequencies of A, U, and G. For mRNA codons that encode amino acids with four synonyms, the trend of base composition is opposite to those encoding amino acids with one to three synonyms; the instability of C could be inhibited or reduced via formation of hydrogen bonds with a G and/or with a protonated C, and the secondary structure of the resultant mRNA could be adjusted via the multiple synonymous alternates at the third position of their codons to facilitate the translation process. The overall pattern of occurrence for C in the genetic code not only minimizes deleterious mutations and favors proper function of the translational machinery by excluding C from certain positions within codons, but also allows the occurrence of genetic diversity via mutation by including C in less-critical positions.

Keywords

Genetic Code, Base Triplet, Synonyms, Cytosine Deamination,

Translation Mutations

1. Introduction

The standard genetic code is nearly universal, and consists of 64 combinations of base triplets made from four different bases—adenine (A), guanine (G), uracil (U), and cytosine (C). Since 61 of the 64 base triplets are used to encode only 20 amino acids, most amino acids are encoded by more than one codon. The remaining three triplets, called stop codons, designate the termination of translation [1]. To the author's knowledge, no study has investigated the pattern of occurrence of cytosine in the genetic code; it thus became the objective of this study. The author explored the base composition and sequence of all 64 codons, and inferred some important features in view of the instability of cytosine.

2. Methods

Since the genetic code is highly degenerate, meaning that most amino acids are encoded by more than one mRNA codon, the author divided the standard genetic codons into two groups: the base triplets encoding amino acids that have one to three synonymous codons (**Table 1**), and those amino acids with four synonymous codons (**Table 2**). Amino acids serine, leucine, and arginine each have six synonymous codons (also called synonyms); they are categorized as two-synonym plus four-synonym occurrences. The author determined the percentage (%) of A, U, G, and C at every position of the base triplet for mRNA codons with one to three synonyms (**Table 1**), and those with four synonyms (**Table 2**), respectively.

3. Results

The first feature is the absence of cytosine (C) in both the start (AUG, also the only codon for methionine) and stop codons (UAA, UAG, and UGA) of translation. The initiation and termination of translation are critical for protein synthesis; therefore, evolution has resulted in a higher frequency of the more stable A, U, and G to avoid a fatal malfunction in the translation process. Cytosine is also absent from the only codon for the amino acid tryptophan (UGG). The author infers that the absence of cytosine from the codons for methionine and tryptophan, neither of which has an alternate mRNA codon, is the result of evolutionary selection to avoid translation errors due to the spontaneous deamination of cytosine to uracil [2] [3] [4].

In contrast to the standard genetic code referred to above, mitochondrial genomes contain alternate start codons (e.g., AUA and AUU in humans, and GUG and UUG in prokaryotes). All vertebrate mitochondria use AGA and AGG as translation terminators. Mitochondrial mRNA from vertebrates and microorganisms use UGA to encode tryptophan rather than as a translation terminator,

Table 1. Analysis of the base triplets that encode the initiation and termination of translation, and those that encode amino acids with one to three synonymous codons. The genetic codons, and the amino acids encoded and their properties are from Berg *et al.* (2015) and Harris *et al.* (2016) [1] [6].

Amino Acid Encoded, Including the Property and Formula of Its Side Chain	mRNA Codon	% of Each Base for mRNA Codons with 1 - 3 Synonyms				
		1st Position (Left)	2nd Position (Middle)	3rd Position (Right)	1st and 2nd Positions	All Three Positions
Methionine (Met) Translation Start Codon hydrophobic $-CH_2CH_2SCH_3$	AUG					
Tryptophan (Trp) hydrophobic $-CH_2C_8H_6N$	UGG					
Lysine (Lys) positively charged $-CH_2CH_2CH_2CH_2NH_3^+$	AAA AAG	% of A 12/32 = 37.5%	% of A 16/32 = 50%	% of A 8/32 = 25%	% of A 28/64 = 43.8%	% of A 36/96 = 37.5%
Asparagine (Asn) polar $-CH_2CONH_2$	AAU AAC	% of U 12/32 = 37.5%	% of U 8/32 = 25%	% of U 8/32 = 25%	% of U 20/64 = 31.2%	% of U 28/96 = 29.2%
Arginine (Arg) positively charged $-CH_2CH_2CH_2NHC(NH_2)_2^+$	AGA AGG	% of G 4/32 = 12.5%	% of G 8/32 = 25%	% of G 8/32 = 25%	% of G 12/64 = 18.8%	% of G 20/96 = 20.8%
Serine (Ser) polar $-CH_2OH$	AGU AGC	% of C 4/32 = 12.5%	% of C 0/32 = 0%	% of C 8/32 = 25%	% of C 4/64 = 6.2%	% of C 12/96 = 12.5%
Tyrosine (Tyr) polar $-CH_2C_6H_4OH$	UAU UAC					
Leucine (Leu) hydrophobic $-CH_2CH(CH_3)_2$	UUA UUG					
Phenylalanine (Phe) hydrophobic $-CH_2C_6H_5$	UUU UUC					
Cysteine (Cys) polar $-CH_2SH$	UGU UGC					
Glutamic Acid (Glu) negatively charged $-CH_2CH_2COO^-$	GAA GAG					
Aspartic Acid (Asp) negatively charged $-CH_2COO^-$	GAU GAC					
Glutamine (Gln) polar $-CH_2CH_2CONH_2$	CAA CAG					
Histidine (His) polar/positively charged $-CH_2C_3H_3N_2$	CAU CAC					
Isoleucine (Ile) hydrophobic $-CH(CH_3)(CH_2CH_3)$	AUA AUU AUC					
Translation Stop Codon	UAA UAG UGA					

Table 2. Analysis of the base triplets that encode amino acids with four synonymous codons. The genetic codons, and the amino acids encoded and their properties are from Berg *et al.* (2015) and Harris *et al.* (2016) [1] [6].

Amino Acid Encoded, Including the Property and Formula of Its Side Chain	mRNA Codon	% of Each Base for mRNA Codons with Four Synonyms				
		1st Position (Left)	2nd Position (Middle)	3rd Position (Right)	1st and 2nd Positions	All Three Positions
Threonine (Thr) polar $-CHCH_3OH$	ACA					
	ACG					
	ACU					
	ACC					
Serine (Ser) polar $-CH_2OH$	UCA					
	UCG					
	UCU					
	UCC					
Valine (Val) hydrophobic $-CH(CH_3)_2$	GUA					
	GUG	% of A 4/32 = 12.5%	% of A 0/32 = 0%	% of A 8/32 = 25%	% of A 4/64 = 6.2%	% of A 12/96 = 12.5%
	GUU					
	GUC					
Glycine (Gly) hydrophobic $-H$	GGA	% of U 4/32 = 12.5%	% of U 8/32 = 25%	% of U 8/32 = 25%	% of U 12/64 = 18.8%	% of U 20/96 = 20.8%
	GGG					
	GGU					
	GGC	% of G 12/32 = 37.5%	% of G 8/32 = 25%	% of G 8/32 = 25%	% of G 20/64 = 31.2%	% of G 28/96 = 29.2%
Alanine (Ala) hydrophobic $-CH_3$	GCA					
	GCG					
	GCU	% of C 12/32 = 37.5%	% of C 16/32 = 50%	% of C 8/32 = 25%	% of C 28/64 = 43.8%	% of C 36/96 = 37.5%
	GCC					
Leucine (Leu) hydrophobic $-CH_2CH(CH_3)_2$	CUA					
	CUG					
	CUU					
	CUC					
Arginine (Arg) positively charged $-CH_2CH_2CH_2NHC(NH_2)_2^+$	CGA					
	CGG					
	CGU					
	CGC					
Proline (Pro) hydrophobic $-CH_2CH_2CH_2-$	CCA					
	CCG					
	CCU					
	CCC					

and vertebrate mitochondria use AUA for methionine rather than for isoleucine [1] [5] [6]. Again, C is absent from these critical codons. While the author will focus on the nucleic genetic code in the following discussion, it is noted that the pattern of occurrence for cytosine seems to be true for mitochondrial codons as well.

The right-hand column in **Table 1** ("All Three Positions" column) provides the total base composition, including total number and percentage of A, U, G, and C in the mRNA codons shown. Overall, A and U residues are more abundant than G and C residues in the codons for amino acids with one to three synonyms. Data presented in **Table 1** ("1st Position" column) provide the base composition at the 5'/left end of the base triplet of the mRNA codons studied. The frequencies of A and U are 37.5% each, whereas G and C residues are less frequent (12.5% each). At the second/middle base of the mRNA codons studied, the frequencies of A, U, and G are 50%, 25%, and 25%, respectively, as shown in

Table 1 ("2nd Position" column). Interestingly, C does not occur at the second position. At the 3'/right end of the base triplet of the mRNA codons studied (Table 1, "3rd Position" column), there is an equal abundance of A, U, G, and C (25% each).

For mRNA codons that encode amino acids with four synonyms, the trend of base composition is opposite to those encoding amino acids with one to three synonyms. As shown in Table 2 ("All Three Positions" column), C and G residues are more abundant than U and A residues for codons encoding amino acids with four synonyms. The frequencies of C and G at the first position of the mRNA codons studied (Table 2, "1st Position" column) are 37.5% each, whereas the frequencies of U and A are 12.5% each. At the second position of the mRNA codons studied (Table 2, "2nd Position" column), the frequencies of C, G, and U are 50%, 25%, and 25%, respectively; A does not occur at the second position. At the third position of the mRNA codons studied (Table 2, "3rd Position" column), there is an equal abundance of A, U, G, and C (25% each).

4. Discussion

Because cytosine is known to undergo spontaneous deamination into uracil, it is evolutionarily favorable to exclude cytosine from codons critical to the initiation or termination of translation. For amino acids that have one to three synonyms, the frequency of occurrence of C in the first and second positions (the root) of their mRNA codons is significantly lower than the frequencies of occurrence of A, U, and G (see Table 1, "1st and 2nd Positions" column). Furthermore, since the middle position of a base triplet is the most critical location for mRNA codon-tRNA anticodon interaction/binding [7] [8] [9] [10] [11], the complete absence of C from the second position that is observed for base triplets encoding amino acids with one to three synonyms is not surprising.

In Table 1, the only mRNA codons containing C in the root are those encoding histidine (CAU and CAC) and glutamine (CAA and CAG). If spontaneous deamination by hydrolysis occurs, histidine will be converted into tyrosine (UAU and UAC), and glutamine will be converted into a stop codon (UAA and UAG). Since histidine and tyrosine both have polar side chains, in theory, this C-to-U mutation may be less likely to introduce significant changes in a protein's structure or function. However, histidine is often found in active sites of enzymes because its imidazole ring-containing side chain is able to perform many different roles in catalysis, whereas tyrosine has a phenol-containing side chain [1] [6]. Therefore, the histidine-to-tyrosine mutation may allow for genetic variation. The C-to-U mutation within a glutamine codon would cause translation to stop. Because humans can synthesize enough glutamine, it is the most abundant nonessential amino acid in the human body; further studies are needed to determine the effects of the conversion of a glutamine codon into a stop codon on human health and on genetic diversity, although the loss of a protein is likely to have deleterious effects.

For amino acids that have four synonyms, the effects of an unstable C on

translation mutations may not be as deleterious as for amino acids with fewer synonyms, due to the high percentages of C and G in the root, and to the existence of multiple synonymous alternates at the third position of these codons. Frederico *et al.* demonstrated that the rate of hydrolytic deamination of cytosine in a double helix was approximately 140-fold slower than in single-stranded DNA at 37˚C [12]; this difference is mainly due to the decreased accessibility of the N3 and C4 positions in a cytosine that is paired to guanine via hydrogen bonds, blocking the attack from water. The mRNA codons encoding amino acids with four synonyms are CG-rich in the root (see **Table 2**, "1st and 2nd Positions" column), which indicates that they have the potential to inhibit or reduce cytosine deamination by folding upon themselves to form a C≡G double helix, and/or to form a hydrogen-bonded C^+-C i-motif if the RNA sequence is C-rich. (Note: Previous studies have proved the existence of i-motifs under physiological pH [13] [14].) Since CG-rich mRNA regions may form complicated secondary structures that hinder the translation process, producing the same amino acid no matter which of the four mRNA bases is in the third position allows the adjustment of the secondary structure of the resultant mRNA.

Table 2 shows that no A is present at the second position of base triplets encoding amino acids with four synonyms. Previous studies have indicated that the second base of mRNA codons determines the hydrophobicity of the encoded amino acids: The majority of codons for hydrophilic (polar and/or charged) amino acids have A in the second position; while the majority of codons for hydrophobic amino acids have U in the second position [7] [15] [16]. From **Table 1**, we can see that hydrophilic amino acids with one to three synonyms have A or G in the second position of their mRNA codons, while hydrophobic amino acids with one to three synonyms have U or G in their second position. From **Table 2**, we can see that hydrophilic amino acids with four synonyms have C or G in the second position of their mRNA codons, while hydrophobic amino acids have U or C or G in their second position. Since the majority of hydrophilic amino acids have two synonyms, it is reasonable that A is absent from the second position of mRNA codons that encode amino acids with four synonyms.

5. Conclusion

In summary, for amino acids that have one to three synonyms, the frequency of occurrence of C in the root of their mRNA codons is significantly lower than the frequencies of A, U, and G. For amino acids that have four synonyms, the instability of C may be inhibited or reduced via the formation of hydrogen bonds with a G and/or with a protonated C. In addition, the "new" secondary structure of the resultant mRNA could be adjusted via the multiple synonymous alternates in the codons' third positions, which could facilitate the translation process. The overall pattern of occurrence for C in the genetic code not only minimizes deleterious mutations and favors proper function of the translational machinery by excluding C from certain positions within codons, but also allows the occurrence of genetic diversity via mutation by including C in less-critical positions. Evolu-

tion is an excellent engineer.

Acknowledgements

This work is supported by the National Science Foundation under Award No. OIA-1458952. Any opinions, findings, and conclusions expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Conflicts of Interest

The author declares that she has no competing financial interests.

Data Availability Statement

All data generated and analyzed during this study are included in this published article.

References

[1] Berg, J.M., Tymoczko, J.L., Gatto Jr., G.J. and Stryer, L. (2015) Biochemistry. 8th Edition, W.H. Freeman & Company, New York, NY.

[2] Nabel, C.S., Manning, S.A. and Kohli, R.M. (2012) The Curious Chemical Biology of Cytosine: Deamination, Methylation, and Oxidation as Modulators of Genomic Potential. *ACS Chemical Biology*, **7**, 20-30. https://doi.org/10.1021/cb2002895

[3] Poole, A., Penny, D. and Sjoberg, B.M. (2001) Confounded Cytosine! Tinkering and the Evolution of DNA. *Nature Reviews Molecular Cell Biology*, **2**, 147-151. https://doi.org/10.1038/35052091

[4] Levy, M. and Miller, S.L. (1998) The Stability of the RNA Bases: Implications for the Origin of Life. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 7933-7938. https://doi.org/10.1073/pnas.95.14.7933

[5] Swire, J., Judson, O.P. and Burt, A. (2005) Mitochondrial Genetic Codes Evolve to Match Amino Acid Requirements of Proteins. *Journal of Molecular Evolution*, **60**, 128-139. https://doi.org/10.1007/s00239-004-0077-9

[6] Harris, D.C. and Lucy, C.A. (2016) Quantitative Chemical Analysis. 9th Edition, W.H. Freeman & Company, New York, NY.

[7] Lehmann, J. and Libchaber, A. (2008) Degeneracy of the Genetic Code and Stability of the Base Pair at the Second Position of the Anticodon. *RNA*, **14**, 1264-1269. https://doi.org/10.1261/rna.1029808

[8] Auffinger, P. and Westhof, E. (2001) An Extended Structural Signature for the tRNA Anticodon Loop. *RNA*, **7**, 334-341. https://doi.org/10.1017/S1355838201002382

[9] Rumer, Y.B. (2016) Translation of "Systematization of Codons in the Genetic Code [I]" by Yu. B. Rumer (1966). *Philosophical Transactions of The Royal Society A*, **374**, Article ID: 20150446. https://doi.org/10.1098/rsta.2015.0446

[10] Rumer, Y.B. (2016) Translation of "Systematization of Codons in the Genetic Code [II]" by Yu. B. Rumer (1968). *Philosophical Transactions of The Royal Society A*, **374**, Article ID: 20150447. https://doi.org/10.1098/rsta.2015.0447

[11] Rumer, Y.B. (2016) Translation of "Systematization of Codons in the Genetic Code [III]" by Yu. B. Rumer (1969). *Philosophical Transactions of The Royal Society A*,

374, Article ID: 20150448. https://doi.org/10.1098/rsta.2015.0448

[12] Frederico, L.A., Kunkel, T.A. and Shaw, B.R. (1990) A Sensitive Genetic Assay for the Detection of Cytosine Deamination: Determination of Rate Constants and the Activation Energy. *Biochemistry*, **29**, 2532-2537. https://doi.org/10.1021/bi00462a015

[13] Wright, E.P., Huppert, J.L. and Waller, Z.A.E. (2017) Identification of Multiple Genomic DNA Sequences Which Form I-motif Structures at Neutral pH. *Nucleic Acids Research*, **45**, 2951-2959. https://doi.org/10.1093/nar/gkx090

[14] Zeraati, M., Langley, D.B., Schofield, P., Moye, A.L., Rouet, R., Hughes, W.E., Bryan, T.M., Dinger, M.E. and Christ, D. (2018) I-motif DNA Structures Are Formed in the Nuclei of Human Cells. *Nature Chemistry*, **10**, 631-637. https://doi.org/10.1038/s41557-018-0046-3

[15] Chiusano, M.L., Alvarez-Valin, F., Di Giulio, M., D'Onofrio, G., Ammirato, G., Colonna, G. and Bernardi, G. (2000) Second Codon Positions of Genes and the Secondary Structures of Proteins. Relationships and Implications for the Origin of the Genetic Code. *Gene*, **261**, 63-69. https://doi.org/10.1016/S0378-1119(00)00521-7

[16] Copley, S.D., Smith, E. and Morowitz, H.J. (2005) A Mechanism for the Association of Amino Acids with Their Codons and the Origin of the Genetic Code. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4442-4447. https://doi.org/10.1073/pnas.0501049102