Journal of Molecular Evolution (2021) 89:589–593
https://doi.org/10.1007/s00239-021-10027-z

LETTER TO THE EDITOR

Codon Usage Bias: An Endless Tale

Andrés Iriarte[1,2] · Guillermo Lamolle[1] · Héctor Musto[1]

Received: 17 June 2021 / Accepted: 6 August 2021 / Published online: 12 August 2021

Abstract
Since the genetic code is degenerate, several codons are translated to the same amino acid. Although these triplets were historically considered to be "synonymous" and therefore expected to be used at rather equal frequencies in all genomes, we now know that this is not the case. Indeed, since several coding sequences were obtained in the late '70s and early '80s in the last century, coming from either the same or different species, it was evident that (a) each genome, taken globally, displayed different codon usage patterns, which means that different genomes display a particular global codon usage table when all genes are considered together, and (b) there is a strong intragenomic diversity: in other words, within a given species the codon usage pattern can (and usually do) differ greatly among genes in the same genome. These different patterns were attributed to two main factors: first, the mutational bias characteristic of each genome, which determines that GC− poor species display a general bias towards A/T codons while the reverse is true for GC− rich species. Second, the differences in codon usage among genes from the same species are due to natural selection acting at the level of translation, in such a way that highly expressed genes tend to use codons that match with the most abundant isoacceptor tRNAs. Thus, these genes are translated at a highest rate, which in turn leads to avoid the limiting factor in translation which is the number of available ribosomes per cell. Although these explanations are still valid, new factors are almost constantly postulated to affect codon usage. In this mini review, we shall try to summarize them.

Introduction

In the standard genetic code, there are 61 sense codons and three stop triplets. Since there are only 20 amino acids, it follows that most of them must be coded by more than one codon. Indeed, while Met and Trp are coded by just one codon, all the other amino acids are coded by two codons (Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, and Cys), three (Ile), four (Val, Pro, Thr, Ala, and Gly) or six triplets (Leu, Ser, and Arg). The different codons that are translated to the same amino acid are called "synonymous codons". Since the availability of the first coding sequences, it became clear that there is a high variability both among and within different species in the usage of the synonymous codons (Grantham et al. 1980; Gouy and Gautier 1982), which is called "codon usage bias" (CUB).

The first ideas about CUB were (a) the genomic GC content of the different species will determine GC3 (that is, the molar content of Guanine plus Cytosine at third codon positions) and therefore codon usage. For example, a GC− rich genome will tend to prefer GC− rich ended codons, while the opposite will be true for a GC− poor species. For clear examples, see the differences in CUB in *Mycobacterium tuberculosis*, a GC− rich species (Andersson and Sharp 1996), against *Plasmodium falciparum*, a very GC− poor species (Musto et al. 1995). These biases in CUB could simply be the result of neutral processes, that is, in simple words, the tendency of the enzymes that replicate and/or repair DNA to incorporate more GC or AT bases. The same would happen in organisms characterized by regions that strongly differ in GC content, like the human genome (Bernardi and Bernardi 1985; Scaiewicz et al. 2006). (b) The second main idea was that intragenomic CUB could be due to certain codons being favored because they increase the efficiency (and accuracy) of translation (for two and

Handling editor: **David Liberles**.

✉ Héctor Musto
hmusto@gmail.com

[1] Laboratorio de Genómica Evolutiva, Depto. de Biología Celular y Molecular, Facultad de Ciencias, Universidad de la República, 11400 Montevideo, Uruguay

[2] Laboratorio de Biología Computacional, Depto. de Desarrollo Biotecnológico, Instituto de Higiene, Facultad de Medicina, Universidad de la República, 11600 Montevideo, Uruguay

classic examples, see Gouy and Gautier 1982; Akashi 1994). This idea was remarkably reinforced by the work of Ikemura, which showed a strong positive correlation between codon usage and tRNA content in several organisms. In other words, he showed that the preferred codons in highly expressed genes matched with the most abundant isoacceptors tRNAs in several species (Ikemura 1985; Kanaya et al. 2001).

Therefore, a kind of "intellectual peace" was predominant in the field of CUB: the codon usage was the result of a balance between biases generated by mutation, natural selection, and random genetic drift (Bulmer 1991). And natural selection was only due to translational speed and accuracy.

"New" Factors that Shape Codon Usage

But we are dealing with biology. And several new layers of complexity on CUB were added in the last years. Let's summarize some of them. Other recent reviews have touched on this subject, where important references can be found (see, for instance, Novoa and Ribas de Pouplana 2012; Novoa et al. 2019; Chaney and Clark 2015; Komar 2016; Hanson and Coller 2018).

(a) CUB is related to the hydrophobicity of the encoded proteins (de Miranda et al. 2000; Romero et al. 2000). In these two papers, using multivariate analyses, it has been shown that in two different Bacteria, among the main factors that shape codon usage, is the hydrophobicity of each protein. We stress that this factor has not been deeply studied, and that preliminary results from our lab indicate that if we compare among prokaryotes the codon usage of transmembrane regions, those that are embedded in the membrane show an increase in G− ended codons in relation to the ones that are located outside the membrane. Currently, we have no explanation for this (preliminary) result.

(b) The location of genes in the leading or lagging strand of replication among prokaryotes (McInerney 1998; Romero et al. 2000). As has been noted in these two papers, in some Bacteria, the location of the genes either in the leading or lagging strand of replication, leads to a clearly differential codon usage. Indeed, those sequences placed in the leading strand display a G and T excess while C and A are predominant at third codon positions. This bias can be so strong, although the reasons remains obscure but is surely related to the different enzymes that replicate each strand, that can even lead to different amino acids content comparing both strands (Lafay et al. 1999). We stress that this bias in codon usage can be seen even in GC− poor species, where translational selection for speed is operative (Musto et al. 2003).

(c) The preference of AGR codons to encode Arg and the avoidance of CGN synonymous triplets in thermophilic prokaryotes (Lynn et al. 2002) and several RNA viruses (Goñi et al. 2012; Moratorio et al. 2013). The reason for this CUB is still unknown. Therefore, the behavior of Arg codons across evolution deserves much attention. Indeed, as has been noted by Novoa et al. (2019) the usage of Arg codons seem to sufficient to cluster species into domains of life. However, we should stress that as mentioned some lines above, (i) the same happens intragenomically in (at least) human genes, and (ii) thermophilic species (which are mainly Archaea) might be "contaminating" Novoa et al. results. For a rather complete review on codon usage in the domain Archaea, see Iriarte et al. (2014).

(d) Among single stranded RNA viruses there is a strong bias against the usage of CpG and UpA dinucleotides, which affects the synonymous codon usage, as reported many years ago by Rima and McFerran (1997).

(e) Among species that are characterized by strongly compositionally heterogeneous genomes (the so called "isochores") mainly, but not exclusively, mammals and birds, genes located in GC− rich regions tend to be GC− rich at third codon positions, while the reverse is true for those genes embedded in GC− poor ones. This result was noted several years ago by simply plotting the GC3 level of each (available at that moment) gene against the GC content of the region where the CDS is placed (Fig. 1) (for original papers, see for instance Bernardi 2000; Musto et al. 1999). Of course, these results are controversial, since currently there is no agreement about either the origin or the evolution of "isochores", and therefore, about the reasons why genes embedded in different regions either increase or decrease their GC3− levels, and as a consequence, their codon usage (Galtier et al. 2018). In other words, it remains unclear if the heterogeneous compositional structure and CUB in mammals and birds is the result of a neutral process (Eyre-Walker and Hurst 2001; Duret and Galtier 2009) or the consequence of natural selection (Bernardi and Bernardi 1986). However, we should stress that a recent paper has shown strong evidence in the sense that natural selection optimizes codon usage in the human genome (Dhindsa et al. 2020).

(f) Furthermore there is a specific pattern of codon usage in the transcriptome of different human tissues (Kames et al. 2020). This result, although not unexpected, has several implications. (i) The global CUB of human (and by extension, of any multicellular species) must be carefully considered, since most tissues display their

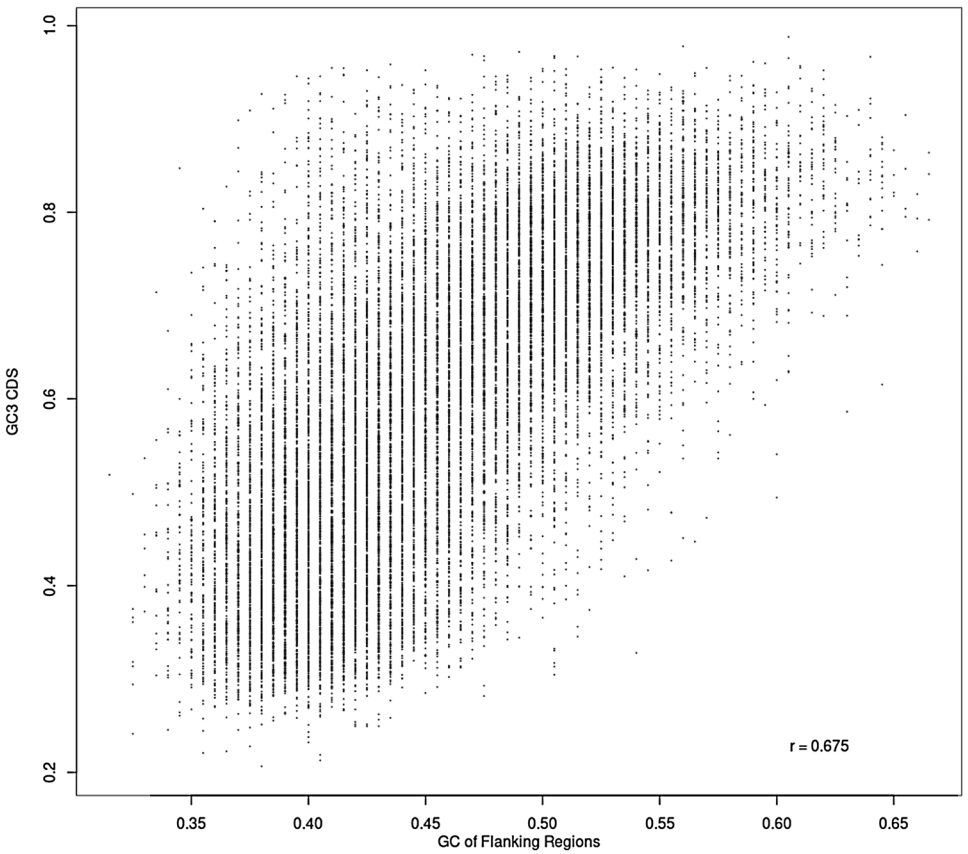

Fig. 1 Compositional correlation between the third codon positions of coding sequences (CDS) and the flanking regions. For each gene the flanking region were defined as 20 kb before the initiation codon plus 20 kb after the stop triplet. The Pearson correlation value, *r*, is shown

own phenotypic codon usage bias. (ii) These authors have shown that CUB is related to the origin of each tissue (ectoderm, mesoderm and endoderm). (iii) This might imply that the concentration of isoacceptor tRNAs can play an important role in the development of multicellular species. (iv) Last, but not least, this differential tissue-specific codon usage might play a role in how different viruses replicate in different tissues, as has been suggested in a recent paper (Simón et al. 2021).

(g) Concerning other vertebrates than mammals and birds, namely *Xenopus laevis* and some fishes, which are more homogeneous from the compositional point of view than mammals and birds, using multivariate analysis, we found that CUB is mainly the result of natural selection acting at the level of speed, that is, that highly expressed sequences use a subset of codons, as happens in prokaryotes. And perhaps more interestingly, several of these "optimal codons" are conserved among these species (Musto et al. 2001; Romero et al. 2003).

But there are more consequences of codon usage. For instance,

(h) Decay of mRNA (Hanson and Coller 2018). In this paper the authors discuss how changes in tRNA pools cause codon-mediated changes in translation speed, which in turn drive to large-scale changes in gene expression, in such a way that lowly expressed mRNAs are degraded faster than highly expressed mRNAs. Needless to say, this might be related to point (e) in this communication.

(i) Protein secretion (Zalucki et al. 2009). These authors reviewed some papers where it is shown that, among prokaryotes, there is a strong usage of non-optimal codons within the signal peptide, and that changing this codons can have deleterious effects on protein folding and export. They suggest that non-optimal codon usage in the signal peptide encodes signals that are important for protein targeting and export to the periplasm.

Of course, some other factors have added complexity to the concept of codon usage. For example, protein folding (Makhoul and Trifonov 2002) and the avoidance of nnUAnn, nnGGnn, nnGnnC, nnCGCn, GUCCnn, CUCCnn, nnCnnA, or UUCGnn dicodons (Tats et al. 2008). Furthermore, CUB affects mRNA splicing (Cartegni et al. 2002); is related to pathologies like cancer (Benisty et al. 2020); changes during the cell cycle (Frenkel-Morgenstern et al.

2012) and so forth and so on. And probably there are many other factors that influence codon usage.

What are the conclusions of this communication? Codon usage appears to be a simple feature, but this is not the case, since as more data are available, it is becoming evident that codon usage goes more beyond than translational selection. Thus, CUB is not the simple addition of all codons in a given species but a very complex trait that needs more in-depth research.

So, we predict that as long as more data becomes available, new layers of complexity will be added to explain CUB; and the "final pattern" observed in a given species and or tissues, results from an equilibrium among different forces, pushing in different directions. Therefore, much more research is needed since CUB appears to be an endless tale in cell biology and evolution. This is the main conclusion of this contribution.

Acknowledgements We thank PEDECIBA and the Sistema Nacional de Investigadores, Uruguay, for partial financial support.

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136(3):927–935

Andersson G, Sharp P (1996) Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology 142(Pt 4):915–925

Benisty H, Weber M, Hernandez-Alias X, Schaefer M, Serrano L (2020) Mutation bias within oncogene families is related to proliferation-specific codon usage. Proc Natl Acad Sci USA 117(48):30848–30856

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241(1):3–17

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22(4):363–365

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24(1–2):1–11

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

Cartegni L, Chew S, Krainer A (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298

Chaney J, Clark P (2015) Roles for synonymous codon usage in protein biogenesis. Annu Rev Biophys 44:143–166

de Miranda AB, Alvarez-Valin F, Jabbari K, Degrave WM, Bernardi G (2000) Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J Mol Evol 1:45–55

Dhindsa R, Copeland B, Mustoe A, Goldstein D (2020) Natural selection shapes codon usage in the human genome. Am J Hum Genet 107(1):83–95

Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genom Hum Genet 10:285–311

Eyre-Walker A, Hurst L (2001) The evolution of isochores. Nat Rev Genet 2(7):549–555

Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, Hou Y-M, Jensen L (2012) Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Mol Syst Biol 8:572

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L (2018) Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol Biol Evol 35(5):1092–1103

Goñi N, Iriarte A, Comas V, Sonora M, Moreno P, Moratorio G, Musto H, Cristina J (2012) Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. Virol J 9:263

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10(22):7055–7074

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8(1):r49–r62

Hanson G, Coller J (2018) Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol 19(1):20–30

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2(1):13–34

Iriarte A, Jara E, Leytón L, Diana L, Musto H (2014) General trends in selectively driven codon usage biases in the domain archaea. J Mol Evol 79(3–4):105–110

Kames J, Alexaki A, Holcomb DD, Santana-Quintero LV, Athey JC, Hamasaki-Katagiri N, Katneni U, Golikov A, Ibla JC, Bar H, Kimchi-Sarfaty C (2020) TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. J Mol Biol 432(11):3369–3378

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, 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(4–5):290–298

Komar A (2016) The Yin and Yang of codon usage. Hum Mol Genet 25(R2):R77–R85

Lafay B, Lloyd A, McLean M, Devine K, Sharp P, Wolfe K (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27(7):1642–1649

Lynn D, Singer G, Hickey D (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 30(19):4272–4277

Makhoul C, Trifonov D (2002) Distribution of rare triplets along mRNA and their relation to protein folding. J Biomol Struct Dyn 20(3):413–420

McInerney J (1998) Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci USA 95(18):10698–10703

Moratorio G, Iriarte A, Moreno P, Musto H, Cristina J (2013) A detailed comparative analysis on the overall codon usage patterns in West Nile virus. Infect Genet Evol 14:396–400

Musto H, Rodriguez-Maseda H, Bernardi G (1995) Compositional properties of nuclear genes from *Plasmodium falciparum*. Gene 152(1):127–132

Musto H, Romero H, Zavala A, Bernardi G (1999) Compositional correlations in the chicken genome. J Mol Evol 49(3):325–329

Musto H, Cruveiller S, D'Onofrio G, Romero H, Bernardi G (2001) Translational selection on codon usage in *Xenopus laevis*. Mol Biol Evol 18(9):1703–1707

Musto H, Romero H, Zavala A (2003) Translational selection is operative for synonymous codon usage in *Clostridium perfringens* and *Clostridium acetobutylicum*. Microbiology 149(Pt 4):855–863

Novoa E, Ribas de Pouplana L (2012) Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet 28(11):574–581

Novoa EM, Jungreis I, Jaillon O, Kellis M (2019) Elucidation of codon usage signatures across the domains of life. Mol Biol Evol 36(10):2328–2339

Rima BK, McFerran NV (1997) Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J Gen Virol 78:2859–2870

Romero H, Zavala A, Musto H (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res 28(10):2084–2090

Romero H, Zavala A, Musto H, Bernardi G (2003) The influence of translational selection on codon usage in fishes from the family Cyprinidae. Gene 317(1–2):141–147

Scaiewicz V, Sabbía V, Piovani R, Musto H (2006) CpG islands are the second main factor shaping codon usage in human genes. Biochem Biophys Res Commun 343(4):1257–1261

Simón D, Cristina J, Musto H (2021) Nucleotide composition and codon usage across viruses and their respective hosts. Front Microbiol 12:646300

Tats A, Tenson T, Remm M (2008) Preferred and avoided codon pairs in three domains of life. BMC Genom 9:463

Zalucki Y, Beacham R, Jennings M (2009) Biased codon usage in signal peptides: a role in protein export. Trends Microbiol 17:146–150