J Mol Evol (2016) 82:245–246
DOI 10.1007/s00239-016-9742-z

RANDOM WALKING

What We Know and What We Should Know About Codon Usage

Héctor Musto[1]

Received: 21 April 2016 / Accepted: 27 April 2016 / Published online: 6 May 2016

In the following short lines, I will try to summarize the meaning of the concept "codon usage," i.e., how different organisms, or different tissues from the same organism, use the 59 codons which code for more than one amino acid (all but AUG and UGG which are the only ones for Met and Trp, respectively, and excluding the stop codons, which are UAA, UAG, and UGA). First of all, let us assume that LUCA had what now we call the "universal genetic code": LUCA had a well-developed translation apparatus, which translated the canonical 20 amino acids, had all the tRNAs, and, of course, all the enzymes needed to charge the tRNAs (aminoacyl tRNA synthetases). What happened before that is still a matter of discussion, which is related more to the origin of the genetic code than to codon usage (Musto 2014).

First of all, before the advent of DNA sequencing techniques, codon usage was postulated as selectively neutral, i.e., given a long enough sequence (or a big number of CDS), each triplet coding for the same amino acid should be equally frequent. That is, for example, given that UUU and UUC code for Phe, both of them should be used at about 50 %. For fourfold degenerate codons, like Pro, CCU, CCA, CCC, and CCG, each of them should be at around 25 % (King and Jukes 1969). However, as long as many DNA sequences were available, it became evident that this was not the case, mainly by the work of Grantham and his colleagues (Grantham et al. 1980, 1981). In their pioneering work, these authors showed two crucial issues: first, each genome apparently had its own strategy for the usage of synonymous codons, and second, genes expressed at higher levels tended to display a different codon usage from the rest of the genes from the same genome. The first point could be easily explained by the different mutational bias (GC-rich organisms tend to use more G- and C-ending triplets, while A- and T-ending codons should be preferred by GC-poor genomes). On the other hand, the second point was by far much more complex and involved more biological implications, since it suggested that in highly expressed sequences, natural selection was at action selecting a subset of triplets. At the moment, that was a revolutionary idea, because its main implication was that third-codon positions were the result of natural selection acting at the level of translation. In other words, synonymous codons, at least in a subset of genes, were neither neutral nor silent. Incidentally, this idea, together with other results, inspired one of the most beautiful titles of a review written on the subject several years later: "DNA sequence evolution: the sounds of silence" (Sharp et al. 1995).

Later, this selectionist hypothesis was confirmed by Ikemura, who showed that in several organisms, the most heavily expressed genes have a tendency to use the codons which match the most frequent tRNAs (Ikemura 1985). His results, both technically and conceptually, were enormous. Indeed, he found not only that the isoacceptor tRNAs (i.e., the different tRNAs which have different anticodons but carry the same amino acid) are not present at the same concentration within a cell, but also that those which recognize (following certain rules) the most frequent codons at the most heavily expressed genes are at the highest concentration. Needless to say, his contribution gave a great impulse to the selectionist theory. Therefore, nothing

✉ Héctor Musto
hmusto@gmail.com

[1] Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias, Departamento de Ecología y Evolución, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay

more simple than accepting that highly expressed genes do prefer certain codons, because they are recognized by the most abundant isoacceptor tRNAs, which naturally leads to the highest speed in translation. This, of course, should lead to "fix" these codons in the heavily expressed genes because ribosomes, the limiting factor in translation, should be readily available to translate other proteins. Apparently, the circle was closed. But of course, another problem arose, a problem which is old in evolution—Which was first: the concentration of tRNAs or the most used codons in highly expressed genes?

Leaving aside this problem, as old as the problem of the egg and the hen, these findings, together with results found mainly by Sharp and Li (1986a, b, 1987), led Bulmer some years later, in his seminal paper based on population genetics, to postulate that the synonymous codon usage among genes in a given species is mainly the result of a balance between biases generated by mutation, natural selection, and random genetic drift (Bulmer 1991).

In my opinion, these briefly summarized papers were crucial. Of course, in these last 25 years, some papers were published which shed more light on this crucial aspect of molecular evolution. For example, the role of accuracy, hydropathy, the location of genes either in the leading or lagging strand of replication, the role of isochores for codon usage, the three-dimensional structure of mRNA, the "dinucleotide effect" when talking of RNA viruses, and several other factors were added to this widely accepted view. (See, for example, Akashi 1994; Romero et al. 2000; D'Onofrio et al. 1991; Musto et al. 1999; Moratorio et al. 2013; and, for a general recent excellent review, see Chaney and Clark 2015).

But the question still remains. Taking into consideration that there are only 59 codons to play with in the universal genetic code, how many characteristics of the genetic code are there and how they influence the rate of translation still remain to be discovered. In my opinion, there are much more physiological features than we know. And perhaps, some subtle combination of them will teach us much more than we guess today about gene expression and regulation. For example, why does ribosomal profiling shows that ribosomes are not mainly "captured" in regions (even with highly expressed genes), "full" of codons which match the most frequent tRNAs, which is something obviously expected from the above-mentioned, and usually accepted, theory? Although this field seems old, it is still fruitful, even more in species like us. However, it is complex by nature. Moreover, the genetic code and its use, in the future, will give us clues to understand the complexity. Hopefully, young colleagues will consider working in this field.

I will finish these reflections with something that happens to me when I think, write papers, or give lessons about this subject. I guess that we perhaps are near the end and that we know all, or nearly all, that must be known, and that we have the complete picture. But a voice comes from far away and whispers slowly, but clearly: "There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy."

References

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

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129(3):897–907

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

D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32(6):504–510

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

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9(1):r43–r74

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

King J, Jukes T (1969) Non-darwinian evolution. Science 164(3881):788–798

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 (2014) How many theories on the genetic code do we need? J Mol Evol 1–2:1

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

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

Sharp PM, Li WH (1986a) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24(1–2):28–38

Sharp PM, Li WH (1986b) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res 14(19):7737–7749

Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4(3):222–230

Sharp P, Averof M, Lloyd A, Matassi G, Peden J (1995) DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci 349(1329):241–247