J Mol Evol (2014) 79:1–2
DOI 10.1007/s00239-014-9638-8

RANDOM WALKING

How Many Theories on the Genetic Code Do We Need?

Héctor Musto

Published online: 14 August 2014

The origin of the genetic code—together with the origin of life—is, with no doubt, the main unsolved mystery in biology. Furthermore, both problems are strongly linked. Indeed, in our opinion, life cannot be understood without the genetic code, and the genetic code is in fact at or near the base of the tree of life itself. So, to understand one of them implies to understand the other. After years of reading several papers on both subjects, we feel that we are too much far away from reaching any clear concept on either of them, or, in other words, we are too distant from arriving to a consensus. So, here, without trying to solve the problem, which is a bit far from our expertise, we just propose the following exercise: let us consider moving backwards in time from now back toward the distant past. In doing this thought experiment, we do not cite any of the myriad papers published on the origin and evolution of the code. This is just because this literature is relatively well known, and our idea is not to make a review, but to try to make a small exercise to go backwards in time. Here, we present 12 points about the code that delineate some of its key features. Although these statements are rife with assumptions, we feel that they can help us to *understand* (i.e., not to *solve*) this problem.

(1) Of course, LUCA had a well-developed translational apparatus, which incorporated the canonical 20 amino acids (AAs), and its genetic code was the one we accept as "universal." Further, LUCA was characterized by possessing the main metabolic pathways that we now know.

H. Musto (✉)
Facultad de Ciencias, Laboratorio de Organización y Evolución del Genoma, Departmento de Ecología y Evolución, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay
e-mail: hmusto@gmail.com

(2) Before LUCA, there existed several organisms with (slightly?) different genetic codes. DNA was their genetic material, but their genetic codes were not yet optimized (as is nearly the one that is "established"). Therefore, because metabolic pathways indeed existed, many errors were made while translating proteins. Consequently, these pre-LUCA organisms were eliminated by natural selection (mainly for errors during the synthesis of AAs, nucleotides, energy production, DNA replication, RNA transcription, translation, etc.).

(3) Before period (2), there were organisms using a code of only two positions in the coding triplet (the ones which now we call the 1st and 2nd). The genetic material was DNA, and the translational apparatus was similar to the one we know.

(4) Prior to period (3), only one nucleotide position (which we know as the 2nd) had a coding property. If this is true, only two kinds of AAs were incorporated: hydrophobic (U2: Leu, Ile, and Val) and non-hydrophobic (any other nucleotide in the 2nd position). From this period moving to the past, genetic material could be either DNA or RNA (or both, in different "organisms"). At this period, stop codons were just U- and A-containing, and Tyr (a rather complex AA) appeared later in evolution, and/or messages ended just because RNA ended. This scenario implies some degree of chemical selectivity in the interactions between nucleobases and AA side chains.

(5) Under this scenario, the main point (under an evolutionary point of view) was to distinguish between these two kinds of amino acids: hydrophobic and non-hydrophobic. This gross division distinguishes between "unfunctionalized" side

chains (i.e, containing only C and H atoms) that usually provide the energy for protein folding and "functionalized" side chains (i.e, containing heteroatoms) that provide reactivity.

(6) Life was evolving in water and, for folding, the proteins' hydropathy was crucial. Furthermore, if membranes were developing (which probably could occur at this time) *trans*-membrane proteins could be important for several and obvious reasons: input of food, recognition among "protocells," external pH, etc.

(7) At this stage, it could become important small AAs, selected for "separate" hydrophobic and non-hydrophobic regions (although not necessarily hydrophilic ones): Gly, Ala, Ser (period (4)), and Thr, all of them coded by G/C 2. This period is crucial because it splits AAs coded by U2 from G/C2, and leaves free A2.

(8) At some time between (4) and (7), A2 is co-opted for coding AAs for protecting DNA (or, most probably, RNA) with positively charged residues (Lys, His, and Tyr).

(9) Again, between (4) and (7), mutational bias (either on RNA or, on a second step, DNA) became important. This became crucial after period (4).

(10) Considering the second nucleotide position, similar AAs became incorporated at some point between periods (3) and (2).

(11) Concerning stop codons, although we have no solid argument (in fact, we have no solid argument for any of the points), we posit that they had been acquired by LUCA, endowing it with a great advantage over other organisms lacking an "end of message." But in our, let us say "guess," given the different codes that *do* exist, they were the latest to be incorporated, mainly UGA.

(12) A crucial point is the development of metabolic pathways to synthesize AAs. We think that they co-evolved with the genetic code between points (2) and (10).

How to apply this to big questions? Ultimately, we only need one theory about the evolution of the genetic code. At present, there are several hypotheses on this issue. However, none of them has been widely accepted and all of them lack robust empirical support. An inherent problem resides in the difficulties in reconstructing ancient events that occurred in environmental conditions that are not completely known. Another difficulty, which is not justifiable, is the lack of efforts to reconcile, instead of opposing, some of the existing hypotheses. For instance, it is possible to imagine a scenario whereby first abiogenically synthesized amino acids captured their cognate codons owing to their respective stereochemical affinities, after which the code expanded according to the coevolution theory, and finally, amino acid assignments were adjusted under selection to minimize the effect of translational misreadings and point mutations on the genome. Such a composite theory is extremely flexible and consequently can "explain" just about anything by optimizing the relative contributions of different processes to fit the structure of the standard code.