J. Braz. Chem. Soc., Vol. 21, No. 6, 1117-1120, 2010.
Printed in Brazil - ©2010 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Article

Prebiotic Chemical Kinetics Imprint on Positional Codon Usage

Ricardo Ferreira,*,[a] Kevin Lai[a] and Roberto D. Lins[b]

[a]*Departamento de Química Fundamental, Universidade Federal de Pernambuco, 50740 Recife-PE, Brazil*

[b]*Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

Foram analisados mais de um milhão de exons pertecentes aos 3 domínios da vida, *Eubacteria*, *Archea* e *Eukarya*. Um número significante de primeiros codons com um excesso de bases idênticas nas suas duas primeiras posições foi encontrado nos genomas das *Eubacterias* e *Archea*. Este excesso tem uma freqüência muito menor nos genomas das *Eukarya*. Propõe-se que esta discrepância depende da cinética de crescimento dos oligoribotídeos na Terra primitiva, que foi diminuindo com o passar do tempo.

Over one million exons across the three Domains of Life have been analyzed and a significant excess of first codons containing identical bases in their first two positions has been found in *Eubacteria* and *Archaea* genomes. This frequency is observed to be much lower in the genome of *Eukarya*. This discrepancy is proposed to depend on the kinetics of oligoribotide/oligoribotide-like growth in the primeval Earth, toned down during further evolution.

Keywords: codon usage, pre-biotic Earth, molecular evolution, data mining, RefSeq

Introduction

Biological evolution, including its molecular and cellular aspects, is nowadays reasonably well understood from its central law – genetic variation followed by natural selection. In contrast, the initial steps of the transition from the chemical (prebiotic) to biological evolution remain a shadowed subject, even though it must have been dictated solely by thermodynamics and kinetics. The transitions from inorganic to living systems was firstly proposed by Oparin in 1924[1] and, independently by Haldane[1] (1928, 1954) in the twenties of the last Century, but we owe to Miller[2] (1953) the first experimental demonstration of the synthesis of bio-molecules, such as amino-acids and ribosides, starting with simple molecules such as H_2O, NH_3, CH_4, N_2, etc. under conditions likely those of the early Earth.[2] Therefore, according to the biogenetic *continuity hypothesis*, biomolecules found in present-day biota are directly connectable to the prebiotic molecules.

In more recent studies,[3-5] guanidine and uracyl and its derivatives were synthesized from simple gases under the probable prebiotic conditions. Among the wide range of supporting evidence available in the literature, we note the work of Ferris and co-workers[6,7] who have conducted a series of studies detecting oligoribotide formation in a 5'-phosphorimidazole solution in the presence of a Na^+-montmorillonite[6] and have synthesized mixtures of oligomers, some of which were 55 monomers long.[7] In a recent experiment, di-and tri-peptides have been synthesized from monochiral amino acid solutions containing COS,[8] a compound that probably was also available in prebiotic environments. Such results allow us to assume that the base sequence of contemporary exons such as those available at the NCBI database[9] might retain vestiges from their purely chemical past.

To that, we shall also assume the reasonable, although unproved, supposition that RNA or RNA-like molecules were the first biopolymers.[10-14] It is well known that the posterior discovery of ribozymes[10,15-17] lent further credibility to a "RNA world" scenario.[12] On the other hand, the selective formation of ribose is rather a difficult process, and remained a matter of debate, even though a number of studies have shown that activated ribonucleotides polymerize to form RNA. Recently, this missing link has been addressed by the work of Powner *et al*.[18] Their experiments have shown that activated pyrimidines ribonucleotides can be formed,

*e-mail: rferreira100@yahoo.com

bypassing free ribose and the nucleobases, starting from cyanamide, cyanoacetylene, glycolaldehyde, glyceraldehyde and inorganic phosphate, all plausible prebiotic feedstock molecules, and synthesis conditions that are consistent with early-Earth geochemical models.

In this paper evidence for the *continuity hypothesis* in the transition from chemical to biological evolution is investigated on the basis of data mining and kinetics of codon build-up from ribotides/ribotide-like molecules.

Results and Discussion

A hypothesis for kinetic imprint on the formation of early exons

In solution (*e.g.*, primeval ocean), ribotide dimerization reactions would consist of condensations between the 3'OH of one ribotide and the 5'HPO_4 of another (or 2'OH→5'HPO_4 in the absence of a catalyst). Hence, a dimerization between A and G can be schematically represented as:

$$5'HPO_4\text{-}A\text{-}(OH)3' + 5'HPO_4\text{-}G\text{-}(OH)3' \rightarrow 5'HPO_4\text{-}A\text{-}G\text{-}(OH)3' + H_2O$$

Such condensation reaction is a function of the relative orientation of the two ribotides during a collision. As a first approximation (disregarding small differences in stereochemistry and electron densities of the 3'-hydroxyl and 5'-phosphoryl groups in distinct monoribotides), it should be the same in all cases, except for identical monoribotides where it should be twice as large. This is because one cannot distinguish the collision $5'HPO_4\text{-}A^{I} + A^{II}\text{-}(OH)3'$ from the collision $5'HPO_4\text{-}A^{II} + A^{I}\text{-}(OH)3'$, whereas, by contrast, the collision $A^{I} + U^{II}$ produces the di-ribotide $5'HPO_4$-AU-(OH)3', whereas the collision $U^{II} + A^{I}$ produces the di-ribotide $5'HPO_4$-UA-(OH)3'. Note that it is not necessary that all ribotides/ribotides-like species had to be present, but only that the concentrations of each monomer to be roughly the same and continuously replenished.

Admitting the 5'→3' growth without questioning its origins here, further polymerization would produce a chain population with an excess of diribotides with identical bases in their first two positions at their 5'-end (*e.g.*, AA, CC, GG, UU – herein referred as XX). A consequence of this simple chemical kinetics effect is that, as the polyribotides were beginning to act as RNAs, the first position of such biopolymers was richer in the above-mentioned doublets. If that would be the case, a higher than the randomic frequency of XX_ codons (ϕ) in the first position of the early genes would be expected, *i.e.*, higher than 16/64 = 0.250, or 16/61 = 0.262, if the evolution-determined three stop codons are excluded (where "_" is A, C, G or U). This ratio should diminish as a result of biological evolution processes, such as genetic drift and natural selection. Although the growth of a polyribotide chain may have occurred by both *n+1* and *n+n* mechanisms, a significant excess of the so-called XX_ codons anywhere in the genome would be less probable. Prebiotic oligoribotide sequences were shorter than the current coding sequences (up to 55-mer)[7] and placement of XX doublets would have to fall into specific positions, *i.e.*, first and second positions of non-starting exons. Ultimately, 3.7+ billion years of selective pressure and adaptation would be expected to wash most (or even all) trace excess of randomly placed XX_ codons by usage needs. A main question lies on whether a prebiotic kinetic imprint would still be present in the beginning of the genes of the current biota.

It is important to note that the synthesis of a polypeptide chain in a living organism is initiated by a particular tRNA, ($tRNA_{(met)}$), in response to a translation initiation codon, almost invariably AUG.[19] Hence, all the peptides begin with methionine during the synthesis, which is post-translationally cleaved. Some bacterial sequences represent a few exceptions to this rule. On the other hand, the model probed here would have taken place before translation and therefore, it is equally important to take this fact into account.

To answer this question, 1,060,114 complete and non-redundant coding sequences across the three domains of life,[20] available at the NCBI Reference Sequence database (15), were parsed and the first (non-AUG) codon in each exon identified (sequence details available in supporting online material). The resulting ratio revealed a remarkable trend shown in Figure 1. The ϕ values for the first position of early organisms (*Eubacteria* and *Archaea*) are higher than a random distribution, even if 26.2% is taken as reference, while this frequency falls near the random index for *Eukarya*'s more complex genomes (Figure 1a). This same trend is not shared by the codon usage of the XX_ codons, *i.e.*, ϕ for any given position of the analyzed genomes (28.2% for *Eubacteria*, 27.8% for *Archaea* and 29.2% for *Eukarya*) (Figure 1a). It shows that the excess of the XX_ codons in the first position of *Eubacteria* and *Archaea* is not a consequence of an overall excess of these codons in those domains. In order to verify the non-randomness of the observed bias in the first position, ϕ values were calculated to several other codon positions and normalized to each domain's usage of the XX_ codons. Interestingly, in *Eubacteria* ϕ progressively diminishes with the distance from the first codon reaching values near its usage of the

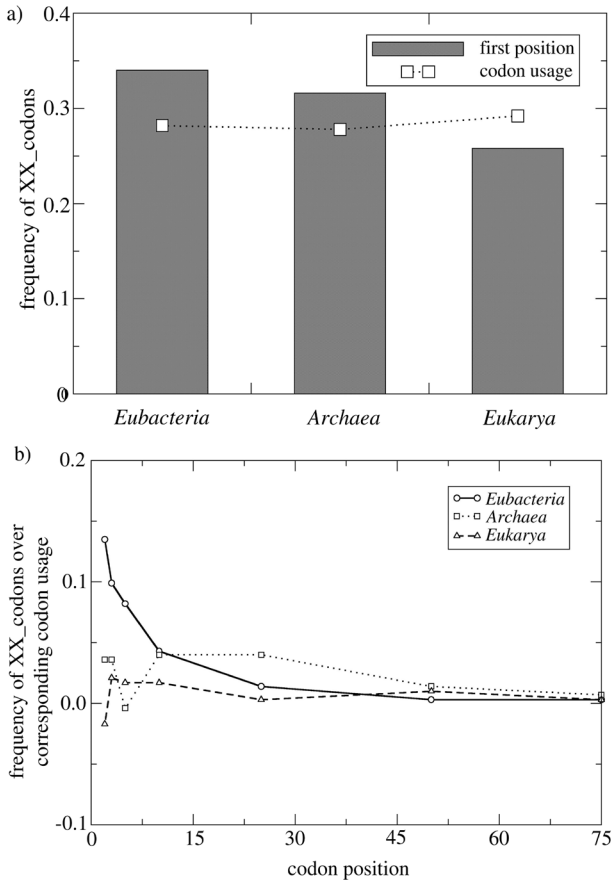

Figure 1. (a) ϕ for the first (non-AUG) codon and the corresponding XX_ codon usage in the three domains of life; (b) ϕ values in arbitrarily chosen positions (2, 3, 5, 10, 25, 50 and 75) normalized to the corresponding XX_ codon usage in each Domain of life.

XX_ codons (within 5%) only after the tenth position (Figure 1b). No significant excess for these codons was detected for *Archaea* beyond the first position as well as, as expected, for *Eukarya*.

Conclusions

The current lack of an evolutionary and/or functional driven explanation for the observed distributions of the XX_ codons in the first position of the genes should not overrule their influence. It is also worth noting that this study employs an elemental statistical treatment and is based on data from modern ribosome-translated reading frames, which do not necessarily correspond to the beginning of prebiotic RNA genes. Nevertheless, considering the odds of evolution and database dependency it is interesting that the analyzed data fits well the above-mentioned hypothesis *per se*. While outlining the transition from the chemical to the biological stage in the prebiotic Earth with a high degree of certainty is hardly possible, this study supports the evidence that the present biota still carries its legacy.

Supplementary Information

Supplementary data are available free of charge at http://jbcs.sbq.org.br, as PDF file.

Acknowledgments

This work was supported in part by the DOE Office of Advanced Scientific Computing Research through the project "Data Intensive Computing for Complex Biological Systems" and the National Council for Scientific and Technological Development, CNPq. The authors acknowledge the use of computational resources provided by the William R. Wiley Environmental Molecular Sciences Laboratory housed at Pacific Northwest National Laboratory. Battelle Memorial Institute operates the Pacific Northwest National Laboratory for the U.S. Department of Energy. We also thank Jason McDermott, Ronald Taylor, Brian Lower, Haluk Resat, Thereza Soares, Victor Russu, Peter Wolynes and André Cavalcanti for providing valuable comments.

References

1. Oparin, A. I.; *The Origin of Life*; Moscow Worker Publisher (in Russian): Moscow, 1924; Haldane, J. B. S.; *On Being the Right Size and Other Essays*, Oxford University Press: Oxford, 1928 (http://www.physlink.com/education/essay_haldane.cfm); Haldane, J. B. S.; *The Biochemistry of Genetics*, George Allen and Unwin Ltd.: London, 1954.
2. Miller, S. L.; *Science* **1953**, *117*, 528.
3. Levy, M.; Miller, S. L.; *J. Mol. Evol.* **1999**, *48*, 631.
4. Levy, M.; Miller, S. L.; Brinton, K.; Bada, J. L.; *Icarus* **2000**, *145*, 609.
5. Levy, M.; Miller, S. L.; Oro, J.; *J. Mol. Evol.* **1999**, *49*, 165.
6. Kawamura, K.; Ferris, J. P.; *J. Am. Chem. Soc.* **1994**, *116*, 7564.
7. Ferris, J. P.; Hill, A. R.; Liu, R. H.; Orgel, L. E.; *Nature* **1996**, *381*, 59.
8. Leman, L. J.; Orgel, L. E.; Ghadiri, M. R.; *J. Am. Chem. Soc.* **2006**, *128*, 20.
9. NCBI - National Center for Biotechnological Information Database; http://www.ncbi.nlm.nib.gov, RefSeq Release 22.
10. Cech, T. R.; *Proc. Natl. Acad. Sci. U. S. A.* **1986**, *83*, 4360.
11. Crick, F. H. C.; *J. Mol. Biol.* **1968**, *38*, 367.
12. Gilbert, W.; *Nature* **1986**, *319*, 618.
13. Nelson, K. E.; Levy, M.; Miller, S. L.; *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 3868.
14. Woese C R.; *The Genetic Code: The Molecular Basis for Genetic Expression*, Harper and Row: New York, 1967.
15. Cech, T. R.; Zaug, A. J.; Grabowski, P. J.; *Cell* **1981**, *27*, 487.
16. Guerriertakada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S.; *Cell* **1983**, *35*, 849.

17. Kole, R.; Baer, M. F.; Stark, B. C.; Altman, S.; *Cell* **1980**, *19*, 881.
18. Powner, M. W.; Gerland, B.; Sutherland, J. D.; *Nature* **2009**, *459*, 239-242.
19. Snustad, D. P.; Simmons, M. J.: Jenkins, J. B.; *Principles of Genetics*, Ed. John Wiley & Sons, Inc.: New York, 1997.
20. Pruitt, K. D.; Tatusova, T.; Maglott, D. R.; *Nucleic Acids Res.* **2005**, *33*, D501.

Received: August 19, 2009
Web Release Date: March 19, 2010

J. Braz. Chem. Soc., Vol. 21, No. 6, S1, 2010.
Printed in Brazil - ©2010 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Prebiotic Chemical Kinetics Imprint on Positional Codon Usage

Ricardo Ferreira,*[,a] Kevin Lai[b] and Roberto D. Lins[a]

[a]*Departamento de Química Fundamental, Universidade Federal de Pernambuco, 50740 Recife-PE, Brazil*

[b]*Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

From the Refseq v22, all complete exon sequences of *Archaea* (28,206) and *Eukarya* (775,406) equal or longer than 75 codons were screened. A randomly chosen subset of the available Eubacterial genome (256,502 containing 75 equal or longer codon sequences) was used. This subset is comprised by all complete and non-redundant sequences of the following species: *Actinobacillus spp., Anabaena spp., Arthrobacter spp., Bacillus spp., Bacteroides spp., Bdellovibrio spp., Bifidobacterium spp., Borrelia spp., Brevibacillus spp., Burkholderia spp., Campylobacter spp., Chlamydia spp., Chlamydophila spp., Clostridium spp., Corynebacterium spp., Deinococcus spp., Desulfitobacterium spp., Desulfovibrio spp., Enterococcus spp., Erwinia spp., Flavobacterium spp., Fusobacterium spp., Geobacillus spp., Geobacter spp., Gordonia spp., Haemophilus spp., Helicobacter spp., Lactobacillus spp., Lactococcus spp., Legionella spp., Leptospira spp., Listeria spp., Mesoplasma spp., Micromonospora spp., Moraxella spp., Mycobacterium spp., Mycoplasma spp., Neisseria spp., Nitrosomonas spp., Nocardia spp., Paracoccus spp., Pasteurella spp., Porphyromonas spp., Prevotella spp., Pseudoalteromonas spp., Pseudomonas spp., Pyrobaculum spp., Rhizobium spp., Rhodobacter spp., Rhodococcus spp., Ruminococcus spp., Selenomonas spp., Shewanella spp., Sphingomonas spp., Spiroplasma spp., Staphylococcus spp., Streptococcus spp., Streptomyces spp., Thermoanaerobacter spp., Treponema spp., Vibrio spp., Weissella spp.*

*e-mail: rferreira100@yahoo.com