Essay

The Genetic Code as a Product of Primordial Viral Evolution

Lev G. Nemchinov *

Posted Date: 26 February 2026

doi: 10.20944/preprints202602.1663.v1

Keywords: genetic code; origin; selfish genetic elements; virus-like structures; protein capsids; evolution

 doi:10.20944/preprints202602.1663.v1

Essay

The Genetic Code as a Product of Primordial Viral Evolution

Lev G. Nemchinov

U.S. Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center, Beltsville, Molecular Plant Pathology Laboratory, 10300 Baltimore Avenue, bldg. 004/115, Mayland, USA, 20705; lev.nemchinov@usda.gov

Abstract

The genetic code is a universal script for life on Earth in which the information is stored in a complex nonbinary base-4 system composed in groups of three. It is generally accepted that during the prebiotic RNA World primordial RNA sequences likely recruited early amino acids using specific number of bases, possibly triplets. There are several hypothetical scenarios regarding how it happened. However, the evolutionary basis and the underlying rationale for codon-specific amino acid assignments are yet to be determined. Among the very first entities on Earth were selfish RNA replicons that lacked a genetic code and could only catalyze their own replication. Supposedly, these primitive selfish elements were essentially unstable naked RNAs that required protective shells to survive. By forming an early protective shell (or capsid) from prebiotic amino acids linked to specific bases, they would not only facilitate the development of a rudimentary genetic code but become precursors to capsid-encoding virus-like structures. This viewpoint suggests that the requirement for genetic material to be encapsidated in a protective protein shell initiated the development of the genetic code, resulting in the formation of the first primitive, virus-like entities preceding the cellular life.

Keywords: genetic code; origin; selfish genetic elements; virus-like structures; protein capsids; evolution

1. Introduction

The genetic code is a universal script for life suggesting that all organisms originated from a single ancestor or "closely interbreeding population" (Crick, 1968). In spite of six decades have elapsed since the best version of the code was presented in 1966 (Cold Spring Harbor Symposium, 1966; Crick, 1968), the exact origin of the code remains a mystery. Several conflicting theories sought to explain the code's origin: the stereochemical hypothesis proposed specific affinity between codons or anticodons and amino acids (Gamow, 1953; Woese, 1965); code adaptation or error-minimization hypothesis proposed that the genetic code adapts to reduce the damage caused by mutations (Sonneborn, 1965; Crick, 1968; Yarus et al. 2005; Koonin and Dolja, 2014); coevolution hypothesis proposed that the prebiotic pathways of the genetic code coevolved with the enzymatic pathways of amino acids biosynthesis (Wong, 1975); and the frozen accident theory suggested that in the "extreme form" codon-to-amino-acid assignments were entirely random and the genetic code is frozen as any change to it would be lethal (Crick, 1968).

While any or all of these hypotheses could be correct, they hardly explain "Why are the codon assignments what they are?" (Koonin and Novozhilov, 2017). In other words, at what point and for what reason did the chemical acids made of carbon, hydrogen, oxygen, nitrogen, and phosphorus atoms transformed into the universal, information-bearing "script" that directs the major processes of life? Was it, perhaps, the fact that primordial genetic elements in the "prebiotic soup" (Oparin, 1924, 1938; Haldane, 1929) were just floating around in a particularly stimulating way that somehow triggered the onset of natural selection process leading to emergence of the genetic code? Or was it

 doi:10.20944/preprints202602.1663.v1

because of the thermodynamics principles influencing the relative abundances of the early amino acids and their correspondence to the composition of the first proteins at the time the genetic code originated (Higgs and Pudritz, 2009)? Or were the early amino acids and perhaps the primitive genetic code or even the first organisms delivered by extraterrestrial sources (Callahan et al. 2011; Crick and Orgel, 1973; Wehbi et al., 2024; Bushman, 2025)? Once again, none of these theories can be bluntly accepted or rejected, some due to the obvious lack of convincing experimental evidence.

Therefore, despite new, important details emerging since the first publications on the origin and structure of the genetic code (Gamow, 1953; Crick, 1968; Woese, 1968), the fundamental question asked by Woese (1968) "…why there exists this particular, unique, precise correspondence between amino acids and codons", remains unanswered. Clearly, though, the question implies that not merely the mechanics and chemistry of the prebiotic interactions between polynucleotides and polyamino acids were crucial for the development of the primitive code and nascent translation system, but also the underlying rationale for all these specific associations. What evolutionary forces could drive these pairings?

2. Main Text

From an evolutionary perspective, among other things, these interactions could be attributed to the proliferation of some of the very first entities on Earth - parasitic genetic elements and selfish replicators (Eigen and Schuster, 1977; Koonin, 2014; Koonin and Dolja, 2014). The earliest parasitic replicons of the RNA World (Gilbert, 1986) unlikely possessed the genetic code and could only catalyze their own replication nonenzymatically, using ribozyme, (Robertson and Joyce, 2012), or with the help of the replicase provided by the putative host RNA system (Furubayashi et al. 2020). Supposedly, they were short, single-stranded RNAs possessing secondary structural elements in the form of several small stem loops (Robertson and Joyce, 2012).

Along with the development of the more sophisticated replication systems, primordial "naked" genetic elements were likely attempting to compartmentalize and secure their molecules inside protective shells by evolving mechanisms to translate RNA. As a result, these capsidless genetic replicons could eventually become predecessors of different classes of viruses, capsid-encoding organisms (Koonin and Dolja, 2014). Given that viral capsids are primarily assembled from multiple copies of a single protein, it is plausible to suggest that the process behind the development of the early genetic code began with the interactions between prebiotic amino acids and bases of the parasitic replicons that were essential to encode this singular protein. In fact, the ability of capsids to self-assemble around the associated nucleic acid is well-known (Cadena-Nava et al. 2012). Although this suggestion is contradictory to the hypothetical origin of viral capsid proteins from cellular ancestors (Krupovic and Koonin, 2017), it is in line with the "virus first" theory (Haldane, 1929; Forterre, 2006) and does not conflict with the definition of viruses as capsid-encoding organisms (Raoult and Forterre, 2008).

Indeed, it is both logical and feasible that the genetic code's origin and evolution were driven by the need to instruct the synthesis of a single, simple protein resembling viral capsid. This assumption positions viruses, via their ancestral genetic replicons, at the very beginning of the genetic code. For instance, the highly conserved, compact jelly-roll capsid protein fold prevalent in non-enveloped icosahedral +RNA viruses infecting all major cellular life forms could be in the group of ancient protein structures derived from the primitive genetic code (Richardson, 1981; Koonin and Dolja, 2014).

Even further, the jelly roll folds could have been preceded by the more primitive, well-conserved small β-barrels protein folds, like Double-Psi Beta-Barrel (DPBB). The DPBB fold is a six-stranded β-barrel that was likely among the first primordial protein structures synthesized by an ancient translational machinery (Yagi et al. 2021; Yagi et al. 2024). Importantly, the DPBB is found in various key enzymes essential for the evolution of the primeval RNA self-replicons, including the core domain of RNA polymerase (Yagi et al. 2021).

 doi:10.20944/preprints202602.1663.v1

The DPBB domain was recently reconstructed using only seven amino acids (Ala, Asp, Glu, Gly, Lys, Arg, Val), demonstrating that this ancient protein structure could have been produced by early translation systems with very limited genetic coding involved, GNN (N = any of the four nucleotides) and ARR (R=A or G), (Yagi et al. 2021). Interestingly, of this set of seven amino acids, at least five (Ala, Asp, Glu, Gly and Val) are considered pre-biotic that is, available abiotically rather than as a result of the genetic coding (Miller, 1952; Bada, 2013; Koonin and Novozhilov, 2017; Longo et al. 2020). Pre-biotic amino acids are thought to be involved in the earliest interactions with the polynucleotides under "primitive earth conditions" (Woese, 1968). Also critical is the apparent nucleic acid-binding ability of the ancestral DPBB (Yagi et al. 2021), as it suggests its possible role in the formation of the ancient genetic code.

Simpler, smaller peptides originated from the primitive code or otherwise minimally-encoded and acting as protective sheath for RNA replicons, could have further evolved by repetition and accretion of their subdomains (Alva et al. 2015; McLachlan et al. 1980), transitioning to more complex capsid-like envelopes preventing RNA from degradation.

Thus, assuming that capsidless selfish replicons were indeed ancestral to all other organisms, their immediate evolutionary advancement would be the formation of a rudimentary genetic code allowing them to produce vital for their survival capsid protein. This would essentially make them capsid-encoding, self-replicating primordial virus-like entities (Koonin and Novozhilov, 2017), potentially serving as precursors to all cellular life forms (Claverie, 2006; Koonin, 2014). Because electrostatic interactions between capsid proteins and RNA are essential for the assembly of modern viruses (Ye at l., 2021), similar forces could have further contributed to the stability of the enveloped virus-like replicons (Blanco et al. 2018).

Although the evolutionary drive to encapsulate parasitic replicons within protein shells might explain the origin and specificity of the first genetic code, which of the tentative mechanisms mentioned above would allow for its formation and evolution in this direction? Regardless of what process took place (Gamow, 1953; Sonneborn, 1965; Crick, 1968; Woese 1965; Wong, 1975), the very first interactions of prebiotic amino acids and code-less genetic elements likely occurred through binding to one another in part due to their abundance and proximity in the "prebiotic milieu" (Woese, 1968). Assuming that the primary role in these interactions belonged to the early replicating oligonucleotides while amino acids were utilized as cofactors supporting primitive ribozymes (Szathmáry, 1999), it cannot be ruled out that recruitment of the prebiotic amino acids had also been selectively directed toward protection and stabilizing of the naked RNA chains.

Subsequently, this early associations between short chains of amino acids and RNA could have created a protective envelope around primitive RNA molecules, where specific amino acids were linked to specific bases, thus facilitating the development of a rudimentary genetic code. Moreover, the first RNA replicons could have been selected for their ability to accumulate amino acids, eventually forming encapsidated RNAs. Indeed "We shall argue that by far the most likely step was that these primitive amino acids spread all over the code until almost all the triplets represented one or other of them" (Crick, 1968).

Given this, a valid question arises about the possible role of the primitive tRNAs (proto-tRNAs) in the origin of the genetic code (Crick, 1968) through ability of the first amino acids to interact with the evolving tRNA anticodons (Giulio, 1998; Hopfield, 1978). The viewpoint proposes that the genetic code initiated by the encapsidation of the early RNA replicons, predated the emergence of tRNA-like minihelices (Bernstein et al. 2016; Lei and Burton, 2021). This is because the evolution of the proto-tRNAs required ribozyme-based mechanisms to generate RNA repeats and inverted repeats (Lei and Burton, 2021) and the primitive aptameric replicons already had various ribozyme activities (Wolf and Koonin, 2007). Furthermore, the viewpoint predicts that the anticodons in the proto-tRNAs may have emerged from the primordial, cognate triplets of the primitive genetic code that was already established in capsid-bearing RNA structures. This aligns with the findings by Rodin et al. (2011) who suggested that primordial tRNAs, tRNA-aminoacylating ribozymes, and the subsequent translation machinery evolved to adapt to the already defined genetic code, rather than the reverse.

 doi:10.20944/preprints202602.1663.v1

It is apparent that primordial encapsidated, ribonucleoprotein structures would be highly reminiscent of archaic virus-like particles, precursors to modern viruses. Therefore, the chain of events involving a) initial binding of prebiotic amino acid to selfish genetic elements to protect and stabilize them; b) development of the rudimentary genetic code as a source of information for the protective envelope; c) establishing specific interactions between amino acids and RNA sequences and forming an envelope around them; and d) emergence of the viral-like particles, offers the natural and most direct explanation for the origin of the genetic code (**Figure 1**).

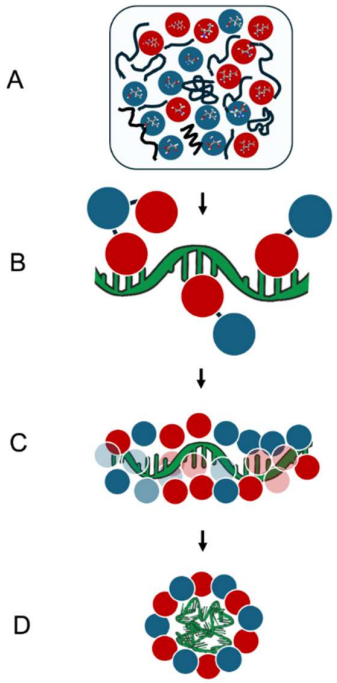

Figure 1. A simplified model for the stepwise emergence of the genetic code. A, initial binding of prebiotic amino acids to primordial genetic replicons in the RNA World environment; B, development of the rudimentary genetic code as a source of information for the protective envelope; C, formation of a protective capsid around RNA replicons; and D, emergence of the viral-like particles. Blue and red circles represent hydrophilic and hydrophobic amino acids; black lines represent primordial genetic elements; oversimplified RNA molecule is shown in green.

Simply put, the need for a protective protein/amino acid coat for early genetic material drove the early development of the genetic code, thus leading to the emergence of the first virus-like entities. This implies that the first information contained in the primitive genetic code to be successfully passed on was associated with the protein shells of archaic virus-like structures.

Setting aside the putative mechanisms allowing the original interactions between RNA replicons and amino acids (Gamow, 1953; Sonneborn, 1965; Woese, 1965; Wong, 1975; Koonin and Novozhilov, 2017; Yarus et al. 2005), this viewpoint focuses on the primary driver, an obvious evolutionary pressure on the selfish genetic elements to adapt by developing the early genetic code for their protective protein shells. Predictably, these first capsid-encoding units transitioned into virus-like forms (Koonin and Dolja, 2014), picking up all components they needed from surrounding environment (Krupovic et al. 2019) to start their assembly lines and eventually paving the way for cellular life forms (Nemchinov, 2025).

3. Conclusions

 doi:10.20944/preprints202602.1663.v1

It is believed that primary amino acids and short peptides likely bound to primitive RNA molecules to stabilize them against degradation (Poole et al. 1998; Noller, 2012; van der Gulik and Speijer, 2015; Pinter et al. 2020). It was also suggested that primordial genetic elements evolved into viruses (Koonin and Dolja, 2014; Krupovic et al. 2019) and that viruses might have preceded the cellular life (Haldane, 1929). Therefore, the two main pillars the viewpoint presented in this article is built upon are not new, having been known and/or proposed before.

What appears to be novel, however, is the alternative interpretation of these theoretical frameworks: early interactions of the prebiotic amino acids with self-replicons could have sparked encapsulation of the primitive RNA to stabilize and protect it and these natural associations prompted the development of the rudimentary genetic code. Thus, the evolutionary force for the emergence of the genetic code was the necessity of forming a protective protein shell around primordial genetic replicons. Accordingly, if the rudimentary genetic code contained instructions for synthesizing protective shells and capsid-encoding replicons evolved into modern viruses, it is accurate to say that the origin of the code was primarily driven by viral evolution. At the very least, the perspective suggests that the origins of the genetic code and viruses can be viewed as inseparable co-evolutionary processes.

Funding: This study was supported by the United States Department of Agriculture, the Agricultural Research Service, CRIS numbers 8042-21500-003-000D.

Informed Consent Statement: The author consents to the publication of the manuscript.

Data Availability Statement: No datasets were generated or analyzed during the current study.

Conflicts of Interest: The author declares that he has no competing interests.

Declaration on the Use of Artificial Intelligence: Generative AI has not been used for writing this manuscript. During the preparation of this work the author used Google Gemini and Google search tools to search and review available literature on the subject, to analyze data as part of the research process, and to improve language and readability. After using the tools, the author thoroughly reviewed and evaluated the content as needed and takes full responsibility for the content of the publication.

References

1. Alva V, Soding J, Lupas AN. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *eLife* 4: e09410.
2. Bada JL. 2013. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chemical Society Reviews* 42: 2186–96.
3. Blanco C, Bayas M, Yan F, Chen IA. 2018. Analysis of evolutionarily independent protein-RNA complexes yields a criterion to evaluate the relevance of prebiotic scenarios. *Curr Biol* 28: 526–537.e5.
4. Bushman FD. 2025. Virolithopanspermia: Might viruses be transported in rocks through space? *PLoS Pathog* 21: e1012955.
5. Cadena-Nava RD, Comas-Garcia M, Garmann RF, Rao ALN, Knobler CM, Gelbart WM. 2012. Self-assembly of viral capsid protein and RNA molecules of different sizes: requirement for a specific high protein/rna mass ratio. *J Virol* 86: 3318–3326.
6. Callahan MP, Smith KE, Cleaves HJ, Dworkin JP. 2011. Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proc Natl Acad Sci* 108: 13995–13998.
7. Claverie JM. 2006. Viruses take center stage in cellular evolution. *Genome Biology* 7: 110.
8. Cold Spring Harbor Symposia on Quantitative Biology. 1966. The Genetic Code. *Cold Spring Harb Symp Quant Biol* 31.
9. Crick FHC, Orgel LE. 1973. Directed Panspermia. *Icarus* 19: 341–346.
10. Eigen M, Schuster P. 1977. A principle of natural self-organization. *Naturwissenschaften* 64: 541–565.
11. Forterre P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* 117: 5–6.

 doi:10.20944/preprints202602.1663.v1

12. Gamow G. 1954. Possible relation between deoxyribonucleic acid and protein structures. *Nature* 173: 318.
13. Gilbert W. 1986. Origin of life: the RNA world. *Nature* 319: 618.
14. Giulio D. 1998. Reflections on the origin of the genetic code: a hypothesis. *J Theor Biol* 191: 191–196.
15. Haldane JBS. 1929. The origin of life. *Ration Annu* 148: 3–10.
16. Higgs PG. 2009. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol Direct* 4: 16.
17. Higgs PG, Pudritz RE. 2009. A Thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* 9: 483–490.
18. Hopfield JJ. 1978. Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. *Proc Natl Acad Sci* 75: 4334–4338.
19. Koonin EV, Dolja VV. 2014. Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiology and Molecular Biology Reviews* 78: 278–303.
20. Koonin EV. 2014. The origins of cellular life. *Antonie van Leeuwenhoek* 106: 27–41.
21. Krupovic M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. *Proc Natl Acad Sci* 114: E2401–E2410.
22. Krupovic M, Dolja VV, Koonin EV. 2019. Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology* 17: 449–458.
23. Lei L, Burton ZF. 2021. Evolution of the genetic code. *Transcription* 12: 28–53.
24. Longo LM, Despotović D, Weil-Ktorza O, Walker MJ, Jabłońska J, Fridmann-Sirkis Y, Varani G, Metanis N, Tawfik DS. 2020. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc Natl Acad Sci* 117: 15731–15739.
25. McLachlan AD, Bloomer AC, Butler PJG. 1980. Structural Repeats and Evolution of Tobacco Mosaic Virus Coat Protein and RNA. *J Mol Biol* 136: 203–224.
26. Miller SL. 1953. Production of amino acids under possible primitive Earth conditions. *Science* 117: 528–529.
27. Noller FH. 2012. Evolution of protein synthesis from an RNA world. *Cold Spring Harb Perspect Biol* 4: a003681.
28. Frenkel-Pinter M, Haynes JW, Mohyeldin AM, Martin C, Sargon AB, Petrov AS, Krishnamurthy R, Hud NV, Williams LD, Leman LJ. 2020. Mutually stabilizing interactions between proto-peptides and RNA. *Nat Commun* 11: 3137.
29. Oparin AI. 1924. The origin of life. *Izd Moskovskii Rabochii* (In Russian).
30. Oparin AI. 1936. *The origin of life.* Macmillan New York, NY.
31. Poole AM, Jeffares DC, Penny D. 1998. The path from the RNA world. *J Mol Evol* 46: 1–17.
32. Raoult D, Forterre P. 2008. Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* 6: 315–319.
33. Richardson ES. 1981. The Anatomy and Taxonomy of Protein Structure. *Advances in Protein Chemistry* 34: 167–339.
34. Robertson MP, Joyce GF. 2012. The Origins of the RNA World. *Cold Spring Harb Perspect Biol* 4: a003608.
35. Root-Bernstein R, Kim Y, Sanjay A, Burton ZF. 2016. tRNA evolution from the proto-tRNA minihelix world. *Transcription* 7: 153–163.
36. Sonneborn TM. 1965. Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications. *Evolving Genes and Proteins* 1: 377–397.
37. Szathmáry E. 1999. The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends in Genetics* 15: 223–229.
38. Furubayashi T, Ueda K, Bansho Y, Motooka D, Nakamura S, Mizuuchi R, Ichihanshi N. 2020. Emergence and diversification of a host-parasite RNA ecosystem through Darwinian evolution. *eLife* 9: e56038.
39. van der Gulik PTS, Speijer D. 2015. How Amino acids and peptides shaped the RNA World. *Life (Basel)* 5: 230–246.
40. Wehbi S, Wheeler A, Morel B, Manepalli N, Minh BQ, Lauretta DS, Masel J. 2024. Order of amino acid recruitment into the genetic code resolved by last universal common ancestor's protein domains. *Proc Natl Acad Sci U S A* 121: e2412152121.
41. Woese CR. 1965. On the evolution of the genetic code. *Proc Natl Acad Sci* 54: 1546–1552.

 doi:10.20944/preprints202602.1663.v1

42. Woese CR. 1968. The fundamental nature of the genetic code: prebiotic interactions between polynucleotides and polyamino acids or their derivatives. *Proc Natl Acad Sci* 59: 110–117.
43. Wolf YI, Koonin EV. 2007. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biology Direct* 2: 14.
44. Wong JTF. 1975. A co-evolution theory of the genetic code. *PNAS* 72: 1909–1912.
45. Yagi S, Tagami S. 2024. An ancestral fold reveals the evolutionary link between RNA polymerase and ribosomal proteins. *Nature Communications* 15: 5938.
46. Yagi S, Padhi AK, Vucinic J, Barbe S, Schiex T, Nakagawa R, Simoncini D, Zhang KYJ, Tagami S. 2021. Seven amino acid types suffice to create the core fold of RNA polymerase. *J Am Chem Soc* 143: 15998–16006.
47. Yarus M, Caporaso JG, Knight R. 2005. Origins of the genetic code: the escaped triplet theory. *Annual Review of Biochemistry* 74: 179–198.
48. Ye L, Ambi UB, Olguin-Nava M, Gribling-Burrer AS, Ahmad S, Bohn P, Weber MM, Smyth RP. 2021. RNA Structures and Their Role in Selective Genome Packaging. *Viruses* 13: 1788.