

# From Matter to Form: the Evolution of the Genetic Code as Semio-Poiesis

Original Study

Suren Zolyan

*Immanuel Kant Baltic Federal University, Kaliningrad, Russia.*

*Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences, Moscow, Russia.*

*Institute of Philosophy, Sociology and Law, National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia*

Received: July 2020; Accepted: September 2020

**Abstract:** *The code is meaningless unless translated.* (Monod 1971, 143)

We address issues of a description of the origin and evolution of the genetic code from the semiotics standpoint. Developing the concept of codepoiesis introduced by M. Barbieri, a new idea of semio-poiesis is proposed. Semio-poiesis, a recursive auto-referential processing of a semiotic system, becomes a form of organization of the bio-world when and while notions of meaning and aiming are introduced into it. The description of the genetic code as a semiotic system (grammar and vocabulary) allows us to apply the method of internal reconstruction to it: on the basis of heterogeneity and irregularity of the current state, to explicate possible previous states and various ways of forming coding and textualization mechanisms. The revealed patterns and irregularities are consistent with hypotheses about the origin and evolution of the genetic code.

**Keywords:** Genetic code, codepoiesis, semio-poiesis, ancestral genetic code, doublet genetic code, grammar of the genetic code

## **INTRODUCTION: ON SEMIOTIC PECULIARITIES OF THE GENETIC CODE: CONTEXT-DEPENDENCY AND ARBITRARINESS**

In this article, we consider the possibilities of the semiotic description of the genetic code. Along with the description of the genetic code's biochemical substance, it is also possible to represent processes as information phenomena and, accordingly, describe them as semiotic systems. This duality can be represented as the dichotomy of biochemical substance and semiotic form if one refers to the cardinal stand on Ferdinand de Saussure's conception: "Language is a form, and not a substance" (1959, 122). Substantial genetic processing characteristics are described as a biochemical interaction, and informational aspects would become the matter of a semiotic description. Having in mind its coding capacity and the textual elements of genetic entities, for some features of the genetic code, the use of semiotic terms seems to be more applicable as they do not have their

correlates in biochemical substance. The dual—biochemical and informational—nature of the genetic code and genome presupposes that one should be based on the principle of complementarity for its description. As in the case of the wave-particle duality of the physical entities, it is impossible to observe and describe both the biochemical and informational properties of genetic entities. Only being taken together, biochemical and informational descriptions will represent a comprehensive state of affairs. The duality of genetic information will be represented through the double theoretical description. We demonstrated it earlier in regard to genetic translation when the biochemical processes can be described in semiotic terms (Zolyan, Zdanov 2018). Unlike other chemical and biochemical phenomena, the genetic information is not some unordered conjunction of elements; it is regulated through their arrangement, linear order, and context. Based on it, the pioneer of bioinformatics in the USSR Vadim Ratner in his last works suggested

## From Matter to Form: the Evolution of the Genetic Code as Semio-Poiesis

to consider the genome not only as a biochemical phenomenon but as a semiotic one also:

“Genes are not the germs of biological structures, but resemble linear texts, written under certain rules and carrying genetic information about molecular structures and functions... Some ‘meaningless’ areas do not encode a protein. In all cases, both genes and non-coding areas are segments of DNA molecules, i.e., are constructed from the same alphabet of four nucleotides. Therefore, the differences between such texts are not in their physical nature, but exclusively in a succession of symbol-monomers. This is the key to the information-linguistic approach. Hence, genes are not physical, but information units of heredity.” (Ratner 2000, 23)

The fact that these characteristics can be formulated as a principle of contextual dependence (sensitivity) was put forward with the emergence of linguistic approaches to the genetic code,<sup>1</sup> although it was ignored by most scholars in favour of context-free formal grammars. The principle means that depending on its location, the same biochemical sequence or entity acquires a different meaning and functional relevance. Contrary to the other chemical or biochemical entities, where an order of components is not essential ( $H_2O$  is equal to  $OH_2$ ), not only a combination of elements, but their arrangement is valid for the genetic coding: AUG is not equal to GAU or UGA. This principle is valid on the higher levels of coding, too: “A protein is like a paragraph written in a twenty-letter language, the exact nature of the protein being determined by the exact order of the letters” (Crick 1981, 48). The same comparison can also be extrapolated regarding genes (operons) and genome. In general, an order of units is a crucial semiotic principle for sign-formation (for example, the words ‘done’ and ‘node’ are composed of the same letters but have different meanings as they have different signified and signifier).

Besides contextual dependence, the arbitrariness of relation between the coding triplets and the coded amino acids is the second principle differentiating the genetic code from biochemical substance and endowing it with the features of semiotic systems. This essential peculiarity had been perpetually mentioned by Francis Crick. The arbitrariness indicated by him is closer to a relationship between signifier and signified. Following the Central dogma, this relation is asymmetric: the nucleotides are “translated” into proteins, but the opposite is impossible, so this is a process of a hierarchical signification. F. Crick considered this arbitrariness the main difference between the genetic code’s regularity from the chemical ones described by the Mendeleev Periodic Table of Elements:

“The Periodic Table would be the same everywhere in the universe. The genetic code appears somewhat arbitrary, or at least partly so. Many attempts have been

made to deduce the relationship between two languages from chemical principles, but none have been successful. The code has a few regular features, but these might be due to chance.” (Crick 1981, 46–47).

Since that time, this assumption of arbitrariness of the genetic code has been repeatedly confirmed despite the still existing prevalence of the opposite “stereo-chemical” point of view:

“It has been shown that there is no deterministic link between codons and amino acids because any codon can be associated with an amino acid. This means that the rules of the genetic code do not descend from chemical necessity and in this sense they are arbitrary. Today, in other words, we have the experimental evidence that the genetic code is a real code, a code that is compatible with the laws of physics and chemistry but is not dictated by them.” (Barbieri 2018, 2).

However, pointing out the genetic code’s arbitrariness is not a solution to the problem, but rather its formulation. Indeed, even in natural language, from ancient times, arbitrariness has been interpreted as the result of certain conventions. Therefore, the question arises of how these conventions have emerged. This question of the origin of semiotic and informational processes led to the issue of the origin of life:

“The origin of life requires understanding the origin of this symbolic control and how inanimate molecules become functional messages. The problem is that conceptually the epistemic cut divides the world in two, and the central problem is how the two worlds are connected. It corresponds to the interpretation that relates the symbol to its referent. In the cell this is an enormously complex process of transcription, translation, synthesis, folding, distribution, and selective control of many proteins. How this coordinated interpreting system originated is the central problem of the origin of life.” (Pattee 2005, 120–121; see also: Chebanov 2019).

The same issue arises concerning the genetic code: which conventions, between whom and by whom, were established to regulate the correspondences between nucleotides and amino acids. In this case, there is no way to refer to some conscious creature, intelligence, or agent:

“There is no chemical affinity that would explain the matching between nucleotide sequences and amino acids. The arbitrary nature of the genetic code was surprising even for its decoders, to the point that Francis Crick, in his speculations about the origins of life, proposed an idea that the rules of code were established by some extraterrestrial intelligence (Crick 1981). As suggested by Crick, the explanations of the arbitrary nature of the genetic code and the mystery of its origins are two interconnected open questions. If the code is arbitrary, in

---

1 Cf.: “In contradistinction to the context freedom of diverse formalized languages, the natural language is context-sensitive, and in particular, its words display a variety of dissimilar contextual meanings. The recent observations on changes in the meaning of codons, depending on their genetic message position, may be noted as a further correspondence between the two patterns.” (Jakobson 1970, 439)

other words, if it is not exclusively dependent on physical or chemical laws but also a kind of codified rules, the only thing that might explain these very rules, is their original codification. The origin of the genetic code represents, however, an unanswered question in biology, precisely like the origin of natural language in linguistics." (Lacková 2018, 309; see also: Lacková, Matlach, Faltýnek 2017).

However, the solution can be found if this parallel with language and linguistics is continued. There are different methods of internal reconstruction; they allow to make certain assumptions about the previous stages of a system based on its synchronic state and its current regularities and irregularities. Thus, there is a possibility to reconstruct the initial stages having the appropriate systemic description for such purpose. The possible description of the genetic code as a semiotic system was presented by Zolyan (2018) and is, on one side, based on certain rules of mapping between two different domains (nucleotids and amino acids), and on a distinction between rules of grammar and vocabulary on the other. Based on it, we intend to develop this approach and apply these semiotic principles of genetic coding to the existing hypotheses on the origin of the genetic code in order to demonstrate how mechanisms emerged.

#### **CODE- POIESIS AND SEMIO-POIESIS AS THE MECHANISM OF EVOLUTION.**

Without infringing on the significance of the experimental data, it makes sense to recall the idea of Jacques Monod: "The third step, according to our hypothesis, was the gradual emergence of teleonomic systems which, around replicative structures, were to construct an organism, a primitive cell" (Monod 1971, 142). After the first two biochemical stages, the new, teleonomic principles should begin to operate, presuming such notions as meaning and purpose. The teleonomic dimension of biosynthesis can no longer be explained on the basis of the characteristics of the biochemical substance. It is in this direction as it seems that the new concept of codepoiesis may be developed:

"The origin of the first cells was based on the ability of the ancestral systems to *generate* the rules of the genetic code, and the subsequent evolution of the cells was based on two complementary processes: one was the *generation* of new organic codes and the other was the *conservation* of the existing ones. Taken together, these two processes are the two sides of a biological phenomenon that can be referred to as '*codepoiesis*'." (Barbieri 2012, 298).

This conception of evolution as a process of codepoiesis and coding disambiguation may be complemented by its consideration as an emergence of semiotic system based on the connection between signified and signifier, that is, the process of semiosis and semio-poiesis. Similarly to indexical relation (*smoke—fire*), the relationship between cause and consequence (nucleotides act as catalysts for the synthesis of amino acids) may

be fixed as a semantic relationship between a signifier (nucleotide sequences, later - codon-anticodon pairs) and the signified (amino acid). Let us remind: "The code is meaningless unless translated." (Monod 1971, 143). Only from the moment of translation, codons and a coding system may emerge. Unlike in the pre-translational period, in RNA codons (or anticodons) no longer directly interact with amino acids—the connection between a signified and a signifier is carried out through special enzymes (ribozymes), and this, even in more degree, resembles a semiotic system. Thus, the evolution of the genetic code shows the embodiment of semiosis, or semio-poiesis: the causal association between accompanying results (indexal semiosis) became a stereo-chemical affinity (iconic correspondence) and, finally, was frozen as an arbitrary correspondence between codons (signifiers) and amino acids (signifieds). This formed an analogue of symbolic signification, and spatial connections between nucleotides (positions) became operational similarly to grammatical categories.

If we take as a ground the no longer contested stand about the arbitrariness of the genetic code, then the search for its biochemical beginnings can be successful only up to a certain limit: namely, before the genetic code arises. The idea of codepoiesis, as it seems to us, makes it possible to deepen the well-known concept of autopoiesis, but it can be expanded. Therefore, the idea of code can be understood not only as a correspondence between two worlds (as it was defined in Barbieri 2012, 2018, 2019a). This correspondence must be expressed in some sign form: a specific dictionary of rules of mappings. But even this seems insufficient, as the concept of code is too narrow for describing teleonomic processes. In addition to the list of correspondences, the notions of meaning (purpose, aim) and grammar (rules for generating new structures from initial symbols) are also required. This statement may be complemented with the idea of a semiotic *scaffolding*:

"Semiosis as an active meaning-seeking-making process often results in the development of some relatively static or even quite solid structures that somehow embed in themselves the findings of active searching-event of semiosis. Thus we may think of semiosis as a process that results in building scaffoldings for further semiosis... Therefore, the evolution of ecosystemic bodies, as systems of scaffolding, presupposes semiosis" (Kull 2015, 231; see also: Hoffmeyer 2007).

Analogies with the simplest codes, such as Morse code, can create erroneous associations. It is not taken into account that Morse code establishes a relationship only between two types of graphic symbols—sequences of dots and dashes, and letters, through which it links itself with a language, its vocabulary and grammar. Dots and dashes then refer to linguistic entities, words, and sentences. Therefore, we believe that it makes sense to take one more step and go from code to sign systems. Codepoiesis then turns out to be an intermediate

link to semio-poiesis—the recursive auto-referential construction of semiotic systems.<sup>2</sup> Since there is no study of such practices, the example of the genetic code can serve not as an illustration, but as a field (a kind of experimental base) for studying this phenomenon.

### SEMIO-POIESIS IN ACTION: THE ORIGIN OF THE GENETIC CODE FROM A SEMIOTIC POINT OF VIEW

Although none of the abovementioned four conceptions of the origin and evolution of the genetic code are not exhaustive, the expressed assumptions allow us to describe, in general terms, the possible patterns of semiosis, or even semio-poiesis. The main semiotic issue is to recognize a correlation between signifier and signified. Only in this case it is possible to speak about signs and meanings. The next step is to establish their functional characteristics and rules of operations with them, transcending from a set of signs to a sign system. Not only the repertoire of correlated elements (nucleotides—amino acids) but the rules for their correlation and combination have to be elucidated. This approach may be identified as a semiotic view on the origin of the genetic code.

The proposed hypotheses can serve as a basis for identifying characteristics of natural semiosis: a transformation of a substance into a sign system. It can be assumed that the initial genetic code was much closer to the biochemical substance; it may coincide with it. The hypotheses put forward generally are consistent with the above-proposed scheme for transition from doublet coding to a triplet one. Almost all versions assume the presence of doublet, and even singlet, coding at the earliest stages of evolution, despite disagreements regarding the functional status of positions. We consider the most influential existing versions in order to demonstrate the possible mechanism of formation of the semiotic interactions.

#### THE FIRST POSITION HYPOTHESES OF THE ORIGIN OF THE GENETIC CODE

In the beginning, in the 90s, one of the most common versions was the one stating that originally, the code was

doublet, and the first position was decisive for mapping:

“The first letter is associated with the steps that turn a simple precursor into an amino acid... Thus, all amino acids formed from the precursor *pyruvate* share the same first letter in the codon... There is a link between hydrothermal vents and the first position of the triplet code. The amino acids can be sorted into a spectrum, running from ‘very hydrophobic’ to ‘very hydrophilic’, and it is this spectrum that bears a relationship with the second position of the triplet code. Five of the six most hydrophobic amino acids have T as the middle base, where as all the most hydrophobic amino acids have have A. The intermediated have a C or G. Overall, then, there are strong deterministic relationships between the first two positions of every codon and the amino acid encoded. If we eliminate the 5 most complex amino acids (leaving 15, plus a stop codons), the patterns in the first two letters of the code become even stronger. It might be, then, that the primordial code was a doublet, and was only later was expanded into a triplet code, by ‘codon capture’; the amino acids competed among themselves for the third position. If so, the earliest amino acids may have had an ‘unfair’ advantage in ‘taking over’ triplet codons, and this seems to be true. The 15 amino acids most likely to have been encoded by the early doublet code hog between them 53 out of 64 triplets... in contrast, the 5 ‘later’ additions muster only 8 codons between them.” (Lane 2009, 72–73).

The characteristic feature of the second position, to distinguish between hydrophilic and hydrophobic amino acids, is related to the environment in which they are formed. As for the first position, one nucleotide in the first position encoded a class of amino acids having a common ancestor: triplets with the same first bases (codon prefixes) encode amino acids with similar biosynthesis pathways:

“It is shown that each of the three codon bases has a general correlation with a different, predictable amino acid property, depending on position within the codon. In addition to the previously recognized link between the mid-base and the hydrophobic-hydrophilic spectrum, we show that, with the exception of G, the first base is generally invariant within a synthetic pathway. G-coded amino acids show a different order, being found only at

---

2 Compare with the other possible differentiation between *code* and *semiosis*: “Single code does not include a source of its change. The code-maker preserves a code; it does not actively change it, because it is not an agent. A new or changed code is a result of learning or choice, or interpretation. However, this does not imply that the code itself carries meaning. Meaning as a phenomenon appears only due to the choice-like (also called abductive) element in semiosis, i.e., in interpretation. Codes are commonly products of interpretation, or semiosis. It is difficult to distinguish between semiosis and code because during the process of habituation, the sign relation gradually loses its choice-making or interpretive aspect, ultimately turning into pure code. This is because the choice in a familiar situation gets solved in an increasingly more predictable way, so the choice becomes hardly noticeable and happens momentarily, before it completely disappears in entirely automatic code... Accepting Barbieri’s definitions of code and arbitrariness, I have shown that code is insufficient for meaning making. The main source of arbitrariness is interpretation, while the arbitrary-relation is afterwards preserved by a code. A code alone is insufficient for semantics. A code is not semiosis” (Kull 2020, 139).

the head of the synthetic pathways. The redundancy of the nature of the third base has a previously unrecognized relationship with molecular weight. The bases U and A (transversions) are associated with the most sharply defined or opposite states in both the first and second position, C somewhat less so or intermediate, and G neutral.” (Tailor, Coates 1989, 177).

This hypothesis was developed in (Copley et al. 2005), where the following reconstruction of the ancient doublet

a codon determines a class of encoded amino acids, the other specifies a member of the class. Such an apparent similarity with the semiotic principles of meaning differentiation naturally gives rise to associations with language: “The apparently systematic nature of these relationships has profound implications for the origin of the genetic code. It appears to be the remains of the first language of the cell, predating the tRNA/ribosome system, persisting with remarkably little change at a deeper level of organiza-



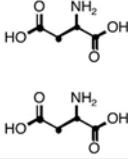

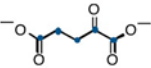
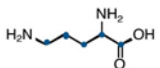

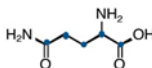
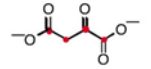
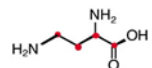
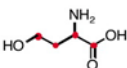
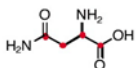
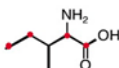

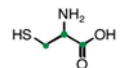
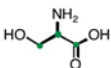
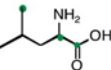
	first position	second position			
		G	C	A	U
	G	Gly	Ala	Asp/Glu	Val
		Gly	Ala	Asp/Glu	Val
					
$\alpha$ -ketoglutarate 	C	Arg	Pro	Gln	Leu
		Orn	Pro	Gln	?
					?
oxaloacetate 	A	Ser / Arg	Thr	Asn	Ile
		Dab	Hsr	Asn	Ile
					
pyruvate 	U	Cys	Ser	Tyr/stop	Leu
		Cys	Ser	?	Leu
				?	

Figure 1: The primordial doublet genetic code—from Copley et al. (2005, 44–46)

code was given—see figure 1:

As follows from the reproduced table, in three cases there is a strong correspondence between the encoding of the reconstructed “progenitor” of the group and its doublet descendant: doublets with cytosine in the first position C1 encode the class of amino acids formed from ketoglutarate, with adenine A1—oxaloacetate and uracil U1—purivate. For amino acids encoded by doublets with guanine in the first position, a common ancestor is not identified, but they also form a particular class. The first position is a kind of a “middle name,” which means belonging to a common ancestor. The second position of the doublet already serves to distinguish between cognate amino acids within the group. If this evolutionary approach is compared with the current situation, it requires to convert the functional relationship between the first and second positions but does not affect the principle itself: one of the positions of

tion than the codon language.” (Tailor, Coates 1989, 177).

The main difference from the current situation is the inversion of the functions of the first and second positions. It can be easily explained as a result of the third position’s appearance as the third position of a codon is coupled with the first position of an anticodon, but is not recognized. Then in order to prevent possible errors during translation, the semantically more critical function is transferred to the second best recognizable position (the second codon position is combined with the second anticodon position). The scheme proposed in Copley et al. (2005) is in perfect agreement with our inner reconstruction of the standard genetic code. It can be considered as a hypothetical doublet phase preceding the emergence of modern quasi- and semi-triplet coding schemes<sup>3</sup>.

On this optimistic note, we could end our survey.

3 The standard genetic code, for the most part, is a superposition of two systems—quasi-triplet (32 codons can be represented as “doublet + non-recognizing third base”); and semi-triplet—30 codons are coded according to the



Unfortunately, or fortunately for the author, in addition to the above-presented version of the reconstructed genetic proto-code, there are several other hypotheses. One of these does not exclude the option that there might be different precursors of the genetic codes, and the diversity of hypotheses of its origin may be a reflection of this possible initial diversity:

“It hardly can be imagined that translation evolved within a single such ensemble. If it emerged on multiple occasions in different ensembles, there is every reason to postulate that there were numerous different codes initially. Why would it be the case that a single code survived? Why was there only one frozen accident (that is, if the actual codon assignments are indeed accidental)?” (Koonin, Novozhilov 2017, 50).

As the reconstructions of the primordial state of affairs have been becoming more detailed, the degree of “grammaticality” of reconstructed coding has increased. This state is described not through purely biochemical or stereo-chemical properties (as presented in Copley et al. (2005), but as being determined by coding rules of composition and operations of distinguishing between signifiers and signifieds. From the point of view of potential code—and semiosis, the next versions seem to be more productive.

### “TWO-LETTERS” THEORIES OF THE ORIGIN OF THE GENETIC CODE.

Most theories of the origin of the genetic code presumed the earliest stage when only two complementary nucleotides were related to amino acids. Two-letter (two-nucleotide) theories can be divided into rigid and non-rigid ones. In the rigid versions, it is assumed that there may be only two nucleotides at the first stage, and usually, cytosine and guanine are mentioned. Under the non-rigid ones, as, for example, in the conceptions of Ikehara and Trifonov, the appearance of adenine and uracil is not excluded, but they are not endowed with meaningful (distinctive and coding) functions.

Initially, it was suggested that the first two nucleotides should have a simpler structure having only two hydrogen bonds, which are uracil from pyrimidine group and adenine (purine). Jiménez-Sánchez (1995) suggested that the genetic code could have begun in an RNA world with these two bases, A and U, grouped in eight triplets coding for the following seven amino acids: lysine, asparagine, tyrosine, methionine, isoleucine, leucine, phenylalanine, and one stop-signal. However, later it was demonstrated that first could appear pairs with three hydrogen bonds: C (cytosine) and G (guanine). Most recent researchers agree on this, although different ways of transforming the ancestral genetic code into the modern state may be a matter for discussion.

As we can see, for the first time, a detailed theory of the two-letter code appears in a series of publications

by Ikehara and his group. According to their concept, the first codons have the form SNS (S and N mean G or C and either of four bases, respectively). SNS structure of the universal genetic code precursor is represented by 16 codons encoding 10 amino acids: Leu, Pro, Val, Ala, His, Gln, Asp, Glu, Arg, and Gly. Repetitive sequences of codons satisfying the SNS formula form the first genes:

“We have concluded that newly-born genes are products of nonstop frames (NSF) on antisense strands of microbial GC-rich genes [GC-NSF(a)] and from SNS repeating sequences [(SNS)*n*] similar to the GC-NSF(a). We have also proposed that the universal genetic code used by most organisms on the earth presently could be derived from a GNC-SNS primitive genetic code. We have reached a novel GNC-SNS genetic code hypothesis. It anticipates a possible evolutionary pathway, suggesting that the universal genetic code has originated from GNC code through SNS code.” (Ikehara 2002, 165; also see Gusev, Schulze-Makuch 2004).

The two-letter concept receives a systemic development in (Frank, Froese 2018). It can be viewed as an implementation of the “rigid” two-letter approach, as at the initial stage it presumed the existence of exclusively G and C. As the authors note: “Various lines of evidence, including the relative prebiotic abundance of the earliest assigned amino acids, the balance of their hydrophobicity, and the higher G.C. content in genome coding regions, indicate that the original two nucleotides were indeed G and C” (Frank, Froese 2018, 259). The authors reconstruct the primordial two-letter (G, C) genetic code, coding four amino acids: CCC: CCG → Pro; CGC, CGG → Arg; GCC; GCG → Ala; GGC; GGG → Gly. As one can see, only the first two positions perform a distinctive function; the third plays only a delimitative role. The authors distinguish three stages of the transition from the primordial two-letter G.C. to the standard genetic code. This process is concretized as follows:

“The first phase includes the primordial two-letter G.C. code. It made it possible to code the four earliest amino acids (Gly, Ala, Pro, Arg) and codons of the standard genetic code that differs from the two-letter code only in the third position. This grouping is consistent because the third position is less reliable, and therefore might not have initially played any coding role. The third position was thus a suitable starting point for introducing new nucleotides. The Phase Two (Gly, Ala, Pro, Arg, Asp, Glu, Val, Leu, His, Gln) adds codons of the standard genetic code that have A or U only in the second position, and those that have A or U both in the second and in the third position. (Gly, Ala, Asp, Glu, Val) all have a first position base G, which they interpret to suggest that the second base was the most important discriminator in the early code. In other words, during this stage the informational content of the code was increased via the addition of a new pair of letters to the second position. The third position continued to accept A and U in combination with

---

principle: “doublet + purine” or “doublet + pyrimidine”. The term *quasi-triplet* was suggested in Frank—Kamenetsky 1980.

the expansion of the second position but, as in stage one, the third position still played no coding role. Phase Three (Gly, Ala, Pro, Arg, Asp, Glu, Val, Leu, His, Gln, Ser, Thr, Cys, Trp, Ile, Lys, Phe, Tyr, Asn, Met) In this final stage of code evolution the first position of a codon also started to include A and U, and this offered new combinations with the other positions" (Frank, Froese 2018, 269).

We would like to draw attention to the fact that these theories demonstrate that certain positions are assigned to specific nucleotides. Thus, the current situation, when any of the three positions of the triplet can be occupied with any of the nucleotides, is to be considered as a result of an evolutionary process.

#### THE HYPOTHESIS OF TWO ALPHABETS.

Eduard Trifonov (2004; 2009) presents the extended version of the two-letter theory. His main emphasis was made on the rules of codon transformation. Most of them are substitutions of cytosine and guanine in the central position for uracil and adenine. Similar substitutions occur in the third position, too. The other type of transformation is the complementary copying; it changes the order of positions: the first becomes the third, the third becomes the first. The initial genetic code includes only two nucleotides G and C, a pair of codons GCC · GGC, and, accordingly, two amino acids (glycine and alanine). Further development is represented as a chain of consequent transformations:

"The next codons to appear should serve valine and aspartic acid, GUC (valine), and GAC (aspartate) are a complementary pair as well as the generic GCC and GGC triplets. The simplest way to derive GUC and GAC triplets from chronologically earlier GCC and GGC is either transition of middle C to U in GCC, with subsequent complementary copying, or transition G to A in GGC, with complementary copying, or both. Next to appear was proline. Today it is encoded by CCC triplet. It should have come simultaneously with its complementary GGG. This codon, in its turn, apparently appeared as mutation of the codon GGC to GGG, in the third redundant codon position. All subsequent codons in the reconstruction below appeared the same way—by change in the third position in one of the earlier codons, and complementary copying" (Trifonov 2004, 7).

Trifonov distributed codons into two groups, belonging into two *alphabets*:

"The complementarity of the simultaneously appearing codons splits the growing amino acid alphabet in two almost independent groups—those amino acids that replace glycine and those replacing alanine. Respective codons are all of the structure N-purine-N in Gly-strand, and N-pyrimidine-N—in Ala-strand. Since all later stages in the codon evolution involve only changes in the third positions and (complementary) in the first positions of the codons, the structure of the codons in the Gly-strand would stay the same, N-purine-N, while N-pyrimidine-N codons are all carried by the Ala-strand.

Hence, two alphabets: G, D, E, R, S, Q, N, H, K, C Y, W - Gly-alphabet, and A, V, P, S, L, T, I, F, M - Ala-alphabet" (Trifonov 2004: 8–9)

According to Trifonov, all the descendants of the two original amino acids preserve common features. As a result of transformations, the derivative history of the first and third positions can be completely changed, but the second position, albeit very slightly, goes back to a common ancestor: all the "descendants" of glycine retain a purine base in the second position, the descendants of alanine retain a pyrimidine one. In accordance with this, in the modern code, Trifonov distinguishes between two alphabets: glycine and alanine. (In linguistic terms, they could be named called purine and pyrimidine *etymologies*, or *derivative histories*). Trifonov suggests that the first genetic messages were written using these two alphabets, which functioned independently for a long time:

"The earliest minigenes, presumably, carried messages encoding miniproteins of two independent alphabets (with the exception of common serine), Gly-alphabet for Gly-strand, and Ala-alphabet for Ala-strand. The most frequent mutations, transitions, that is purine-to-purine and pyrimidine-to-pyrimidine replacements, keep the N-purine-N and N-pyrimidine-N triplet structures unchanged, no matter what happens to the 1st and 3rd positions. In other words, the ancient binary pattern should have stayed rather conserved, at least for some time. The replacements almost exclusively occur within the respective two alphabets, the swaps between the alphabets being rather rare" (Trifonov 2004, 8).

Trifonov's results may also be compared with what we have called cytosine and adenine coding types, or quasi-triplet, and semi-triplet coding. Another observation may relate to the function of the second position in the process of derivation: our synchronic formal model of triplet generating in this theory is represented as a diachronic pattern of evolution.

#### THE FOUR COLUMNS THEORY: STAGES OF CODIFICATION.

The other vision of the primordial genetic code and its evolution is suggested in the so-called theory of four columns (Higgs 2009). It is based on the following assumptions: (i) the earliest amino acids in the code were easiest to synthesize non-biologically, namely Gly, Ala, Asp, Glu, and Val; (ii) these amino acids are assigned to codons with G in the first position—therefore, the first code may have used only these codons; (iii) the code rapidly developed into a four-column code where all the codons in the same column coded for the same amino acid: NUN = Val, NCN = Ala, NAN = Asp and/or Glu, and NGN = Gly; (iv) later amino acids were added sequentially to the code by the process of subdivision of codon blocks in which a subset of the codons assigned to an early amino acid was reassigned to a later amino acid; (v) later amino acids were added into positions formerly occupied by amino

## From Matter to Form: the Evolution of the Genetic Code as Semio-Poiesis

acids with similar properties because this can occur with minimal disruption to the proteins already encoded by

	U	C	A	G	
U	Val	Ala	Asp	Gly	U C
C					U C
A					A G
G					U C A G

Fig. 2: Proposed four-column structure of the earliest genetic code. (Higgs 2009, fig. 2)

the earlier code. As a result, the properties of the amino acids in the final code retain a four-column pattern that is a relic of the earliest stages of code evolution.

Like in other conceptions, the following ten amino acids are distinguished as the original ones: firstly, Gly, Ala, Asp, Glu, Val, and then Ser, Ile, Leu, Pro, Thr. The following stages of the evolution were suggested:

“It is remarkable that the top 5 amino acids on our list (i.e., those with highest prebiotic concentrations) are precisely those that occupy codons with G at first position. This leads us to propose a very early version of the code that used only these GNN codons. However, if three-quarters of the codons were unassigned, all mutations occurring at 1<sup>st</sup> position would render the gene non-functional or impossible to translate. The four-column code is a triplet code, although only the middle base in the codon specifies any information. However, it is possible to code for only four amino acids. Learning to distinguish purines from pyrimidines at the 3<sup>rd</sup> anticodon position creates two blocks of 8 codons in a column, while learning to distinguish purines from pyrimidines at the 1<sup>st</sup> anticodon position creates 8 blocks of 2 codons.” (Higgs 2009, 10).

As one can see, the theory of four columns describes the evolution of distinctive properties of a particular position—at first, only the second position possesses this ability, then the first and the second. As for the third position, its faculty to recognize and distinguish was developed and manifested only partially—it “learned” to distinguish purines from pyrimidines. It remains unclear how the jump from one-element code GNN to four elements (four columns) code NAN, NGN; NCN; NUN had happened. However, in this scheme one can also see a transition from (i) a rigid causal connection between a position and nucleotides (G is obligatory for the first position, the other positions are out of recognition) to (ii) a coding connection between position and function:

only the second position became functionally relevant and any of the 4 nucleotides may occupy it; position, in this case, acts as a formal category regulating characteristics and functioning of an element. Thus, the steps of the code- and semio-poiesis may be identified as a process of refinement and autonomization of grammar and vocabulary. This also led to differentiation within elements (nucleotides—four instead of one) and categories (three recognizing positions instead one). As a reflex of a four-column situation, one can refer to the crucial role of the second position—the second column may be identified with it. As for the first position, in the modern genetic code the primordial codon GNN may be recognised in its modern modification as being assigned for the oldest non-biologically originated amino acids: GCN → Ala; GGN → Gly; GUN → Val; GAY → Asp; GAR → Glu.

### FROM SINGLET TO TRIPLET:

#### A PREFIXAL-SUFFIXAL CONCEPT.

As is evident from its key terms, the most linguistic concept is the one outlined in (Wu et al. 2005). The authors explicitly referred to Crick’s speculation about the possibility of a triplet code, while only one letter is recognized. They call this code a singlet. On the next stage, two letters were begun to be recognized, then three, and it leads, respectively, to doublet and triplet coding. The authors associate the second position with the root, the first with the prefix, and the third with the suffix. Such analogies were not new (the same parallels were drowned in (Rumer 1966), but in this case, suffixes and prefixes are considered as separate paths of the evolution of the genetic code. This strengthens linguistic associations, esp. having in mind that in linguistics, suffixes and prefixes are considered not only as components of a word but also as special ways of word-formation. According to this hypothesis, the code evolution involved a progression from singlet to doublet to triplet codons with a reading mechanism that moves three bases each step:

“We suggest that triplet codons gradually evolved from two types of ambiguous doublet codons, those in which the first two bases of each three-base window were read (\_prefix\_\_ codons) and those in which the last two bases of each window were read (\_suffix\_\_ codons). The evolution of triplet codons from singlet codons via two types of doublet codons. Illustrated is the hypothesis that the current triplet genetic code originates from a singlet code in which the reading mechanism moves three bases per step, but only the middle base of each codon specifies the encoded amino acid. The subsequent doublet code is comprised of two types of codons: prefix codons, in which the first two bases are read (as suggested by Crick), labelled x\_\_; and suffix codons, in which the last two bases are read, labelled \_x. In the doublet-to-triplet codon expansion, prefix doublet codons (x\_\_) evolve to triplet codons in which the third base is read for the first time, and suffix doublet codons (\_x) evolve to triplet codons” (Wu et al. 2005, 55–56).



A question may arise: if one accepts that two codes, the suffixal and the prefixal ones, functioned simultaneously, how to explain the difference between the first and third positions—in one case, the original prefixal coding requires a non-distinguishing nucleotide ending, which is completely consistent with the situation of 32 four-fold codons in the modern standard genetic code. But in the second, suffixal case, the doublet might be transformed into triplet due to the non-distinguishing nucleotide prefix, that is, the first position. This situation is inconsistent with the current situation. Perhaps anticipating such questions, the authors offer the following explanation (*Ibid*, 58–60):

“Initially the first and third bases of the triplet could have played a role in stabilizing the anticodon–codon interaction before they were utilized for encoding an expanding number of amino acids. Is there a reason to believe that a primordial code had some intrinsic capacity to define a three-base reading frame with a central encoding nucleotide? It has been postulated that the central base of the anticodon triplet originates from the second position base of the primordial acceptor minihelix (Rodin et al. 1996; Schimmel 1996) and this second base complementarity appears to be conserved in contemporary tRNA. The acceptor–anticodon complementarity relationship for the second base appears to be a vestige of an ancestral genetic code and is in support of the encoding central base proposed in this model. The doublet code model, using prefix and suffix codons, presents a possible historical explanation for the combinations of fourfold and two-fold degenerate codons. The four-fold degenerate triplet codons in our model originate from prefix codons and the two-fold degenerate triplet codons originate from either prefix or suffix codons in the doublet code” (*Ibid*, 62–63).

In this regard, it makes sense to recall the special role of the triplet composition and its advantages over dinucleotides and other oligonucleotides.<sup>4</sup>

## DISCUSSION

As was mentioned in the above-cited reviews, there is still no commonly accepted general view on the genesis and evolution of the genetic code, and various evolutionary mechanisms are a matter of discussion. However, many of the put-forward ideas have significant explanatory power. A remarkable change in emphasis occurred:

discussions on stereo- and biochemical characteristics of nucleotides and amino acids gave a way to an analysis of functional and structural-systemic characteristics of the coding process itself. At the same time, despite essential differences and incompatibilities, a common core has been formed. It makes it possible to speak with sufficient certainty about the most important characteristics of the primordial genetic code(s):

“The 2-1-3 model and similar schemes, inspired by the properties of the SGC postulate that, in the ancestral code, only the second base of the codon was informative, and this code expanded by assigning specificity to the first, and then—in some codon series—the third bases. The ambiguity reduction model postulates that in the ancestral code, codon series ambiguously

encoded groups of amino acids, such that its subsequent evolution involved a gradual increase in the specificity of codon–amino acid mapping. Postulating that the third letter of the codons was uninformative in the primordial code and reducing the code table to the codons for the ten early amino acids resulted in the reconstruction of the ancient doublet code, albeit with several codon series remaining ambiguous” (Koonin, Novozhilov 2017, 53–54).

Crick’s intuition allows him to formulate the general frame in a few sentences. It became a subject of specifying and detailing in succeeding hypotheses; this was especially salient when manifested in the four-column and prefix-suffix theories. The following observations may be accepted as points of the consensus. Almost all concepts consider the second position as pivotal for coding. Another aspect of consensus is the assumption that the complementary pair with three hydrogen bonds (guanine–cytosine) appeared first, and only then—pairs with two bonds (adenine–uracil). The repertoire of the original amino acids was specified. The first singlets (cytosine or guanine) and doublets were also identified: most probably, it was a dinucleotide pair of guanine and cytosine. Based on it, eight doublet codons may be formed. In this case, singlets and doublets should be inserted in a triplet frame: for special reasons, dinucleotide and oligonucleotide sequences were fixed in the form of trinucleotide sequences, and finally, they were shaped as a triplet. Numerous variations were possible. For example, in the trinucleotide, only the cytosine or guanine in the central position might be relevant, while the first and third positions could be occupied by any of

4 Cf.: “Thus an advantage of trinucleotide over single nucleotide addition is that trinucleotides H-bond longer to the RNA template, giving our hypothetical replicase more time for polymerization. For example, we expect the trinucleotide AGU to base pair longer to its complement UCT, than simply a U opposite a T. Increasing the length of an oligonucleotide increases the stability of base pairing, which can become unexpectedly stable in the interaction of two complementary triplets of tRNAs (Grosjean et al. 1976). However, the number of possible oligonucleotides also increases exponentially; there are 16 pairs of nucleotides (A.A., A.C., AG, etc.) but 64 possible triples (AAA, AAC, etc.)—for each additional nucleotide there are four times as many potential substrates. In principle, oligonucleotides larger than triplets are possible, but only triplets are consistent with the origin of the triplet code for protein synthesis. This is really a description of the problem, and does not explain how the length of the code (a triplet) could have arisen” (Penny 2005, 652; see also: Patel 2005; Štambuk et al. 2018).

the nucleotides: NGN, NCN. Thus, the singlet G or C itself forms its left and right periphery, which may be considered as a precondition for the linear ordering of the genetic information. At the same time, it can be assumed that at the initial stage, the recognition process began outside such a framework, so only the first nucleotide in a non-fixed sequence turned out to be relevant (for example, G; GN.; GNN; GNNNN, etc.).<sup>5</sup> Of course, such freedom critically reduced the possibilities of differentiation of amino acids, and, accordingly, proteins. Subsequently, for the coding of cognate amino acids, differentiation and assignment of degenerate codons for new amino acids occurs. (In semiotic terms, this is a transition from synonymy, interchangeability, to antonymy, mutual exclusion). In all theories, especially those based on the concept of coevolution, the synthesis of cognate amino acids is accompanied by the transformations of the codon encoding their “parent” amino acid.<sup>6</sup>

One can also find a certain correlation between the type of code (two-element or four-element) and coding units (doublet and triplet): two-element codes operated with doublet, four-element code generates triplets and three-letter reading frame.

### CONCLUSION

Considering possible ways of the origin and evolution of the genetic code, we attempt to describe it as a process of semiosis: i.e., as emergence of sign relations, regulations, and operations. The primordial stereo- and biochemically motivated association of nucleotides and amino acids was replaced by arbitrary correspondence. By that time, due to its arbitrariness, these correspondences could not be changed within the code in question. The genetic code emerged as a system of formal restrictions imposed on a biochemical substance. As was already mentioned by Barbieri,

“The statement that the genetic code is a set of constraints is formally correct because its rules impose severe limitations on a virtually unlimited number of possibilities. It must be underlined, however, that they are not physical constraints... The rules of the genetic code are biologically generated constraints that in no way can be assimilated to physical constraints because the genes of the molecules that implement the code are constantly subject, like all other genes, to mutation and neutral drift” (Barbieri 2019, 14).

These abovementioned conceptions create an

opportunity to presume the following stages of this process:

1. Amino acids appear as a by-product of the synthesis of nucleotides, first of guanine and cytosine, and later—of adenine and uracil. The appearance of doublets is the development of a linkage between complementary nucleotides: G–C, A–U. Dinucleotide combinations are precursors of the doublet code. At the first stage, the association between nucleotides and amino acids is mainly of stereochemical affinity.

2. Dinucleotides in the oligonucleotide chain are marked by an adjacent nucleotide, which leads to the creation of a trinucleotide, later—to triplet. The third nucleotide separates coding nucleotide(s) from others. Functionally, this trinucleotide operates as a singlet. Accordingly, the function of recognition is shifted from the first nucleotide that is recognized, but the second (central) one.

3. Initially, only inorganic amino acids were present. The next five were synthesized on the basis of the first five, and then also ten organic amino acids.

4. The chronology of the emergence of amino acids can be represented as transformations of the original degenerate codons. The transformation is going in two directions:

- a) positions are endowed with the capacity of recognition—first to distinguish purines from pyrimidines, then within these groups: cytosine from uracil; guanine from adenine. Due to such differentiation, di- and trinucleotide sequences acquire the features of semiotic structures—doublets and triplets, and positions become functionally different categories;

- b) the composition of codons becomes fixed, substitutions and permutations occur within the codons, allowing to convert previously synonymous codons (that is, encoding the same amino acid) into antonymous (encoding differing, but cognate amino acids).

At the same time, despite the fact that almost all theories contain references to nonsenses, it is noteworthy that there is no explanation for their emergence. Perhaps this is because none of the concepts considers the principles of textuality, as well as contextual dependence when the same codon can be interpreted differently depending on its environment (as in the case of methionine and start codon in the standard code, and stop-codons and non-standard amino acids selenocysteine and pyrrolysine in non-standard codes). Such a situation is hardly accidental—the principles of textuality and contextual dependence can rarely be

5 Cf.: “Coding triplets in this primal pre-translational code were likely similar to the anticodons, with second and third nucleotides being more important than the less specific first one. Later, when the code was expanding in coevolution with the translation apparatus, the importance of 2-3 nucleotides of coding triplets ‘transferred’ to the 1-2 nucleotides of their complements, thus distinguishing anticodons from codons” (Rodin et al., 2011, 6).

6 Cf.: “Under this coevolution theory, the code evolved by subdivision: In the ancestral code, large blocks of codons encoded the same amino acid but were split to encode two amino acids upon the evolution of the respective metabolic pathways. The specific pattern of codon reassignment is determined by the precursor–product relationships between amino acids, whereby a product takes over some of the codons in the block that initially encoded the precursor” (Koonin, Novizhilov 2017, 52).

explained as biochemical phenomena; they have a lingua-semiotic and textual nature. While speaking about the reassignment of coding functions when synonymous codons begin to code for new amino acids, the authors of the abovementioned concepts do not refer to the conditions of such redistribution. Meanwhile, in relation to the current situation with methionine, and non-canonical selenocysteine and pyrrolysine, it is known that such recoding requires special contextual conditions. As for nonsenses, the very appearance of codons marking the beginning and the end of a message, maybe its explanation is to be formulated in terms of the communication theory and text semiotics. Nonsenses, like triplets, are a manifestation of the transition from an isolated natural sign-index semiosis (smoke is a sign of fire, a nucleotide is a catalyst of an amino acid) to systemic semiotics, using sign elements for coding of its internal operations. Nonsenses represent a special type of signs, like conjunctions or meta-names in natural languages: nothing corresponds to them out of the code where they are used, and their meaning coincides with rules of their usage.

As for the distinction and delimitation of triplets, this issue cannot be considered cleared up. Obviously, a transition from oligonucleotides to a fixed structure should have taken place; this implies a delimitation of meaningful segments and prevents triplets' superposition; both are aimed at reducing ambiguity and minimizing errors. Although the very concept of error presupposes the existence of some norms, the possibility of preliminary norm-making processes is not to be excluded. Perhaps, in this case, the principle of sub-textualization also operates, separating the minimal significant units of a text. In any case, it is evident that oligonucleotide sequences had to obtain some textual features: from this point of view, a triplet is the smallest meaningful unit of a message. A message is a sequence of triplets between start- and stop- codons.

The problem of the *reading* frame's emergence is similar: the term itself indicates its semiotic, not biochemical nature. It could perhaps be explained as a result of linearization of the genetic code, when as an initial step, relationships in time (*before—after*) began to be fixed as a linear relationship from left to right. Accordingly, preceding and following non-functional (un-readable) nucleotides began to act as left and right benchmarks around the recognizing nucleotide, and in the end this led to the appearance of a rigid three-element reading frame (centre; left and right boundaries).

Probably, the difficulties of biochemical explanation for the emergence of nonsenses and reading frame are fundamental and even self-contradictory. Before the

genetic translation, there is no way to speculate about its mechanisms, and only its preconditions may be considered: "The code cannot be translated otherwise than by-products of translation" (Monod 1974, 143). Accordingly, no explanation can be given otherwise than by terms of translation. However, in such case it is possible to consider the preconditions of translation. Then the process can be assumed as a transition from a stereochemical and biochemical substance to a form—i.e., as a formation of vocabulary and a grammar. Code formation follows the way of reducing ambiguity, and the formation of grammar is directed by assigning functions to positions. Naturally, these processes require new teleonomic notions deriving from conceptions of code- and semio-poiesis to be described and explained.

As the substance becomes more complicated, biochemical regularities are supplemented by (or even give way to) linguistic and semiotic principles. A biochemical substance no longer determines to code, and relations arise that resemble characteristics not so much for biological systems as semiotic systems (arbitrariness, contextual dependence, linear order, semantic relations, textual characteristics). In addition to the canonical, there are 23 variants (dialects) of the genetic code<sup>7</sup>. The genetic code is not something eternally and invariably existing ("Language of God, the language of life"), but is the product of multi-stage evolution, which leads to the appearance of various synchronic and diachronic variants of the genetic code. As it becomes more complicated, the mechanisms of its organization become closer to the principles of natural language organization.

The evolution of the genetic code can be viewed as a process of semio-poiesis—semiosis in action. The genetic code was born out of the matter, just like the organic world grows out of the inorganic by introducing new organization principles of autopoiesis. Thus, the association of material phenomena (in this case, nucleotides and amino acids) led to creation of semiotic connections. As a final result of random processes, mechanisms for storing and transmitting information emerged, providing the possibility for stable forms of life. The increasing complexity of the organization leads to the crystallization of informational and semiotic principles. Semio-poiesis, a recursive auto-referential processing of semiotic system, becomes a form of organization of the bio-world when and while notions of meaning and aiming are introduced into it.

#### AKNOWLEDGMENT

This research is supported by the grant from the Russian Science Foundation (RSF), project no. 17-18-01536

7 Cf.: "More than 20 nonstandard codes have been described, and new variants continue to emerge with the progress of genomic and metagenomic sequencing. Modifications to the code belong to three major categories: (a) reassignment of codons within the canonical set of 21, including the stop signal; (b) loss (unassignment) of codons, and (c) incorporation of new amino acids. Stop codons are strongly overrepresented among the code modifications. Of the 23 nonstandard codes surveyed by Sengupta & Higgs, there are 8 cases of stop codons being reassigned or acquired, 8 cases of codon loss, and 10 reassignments of a codon from one amino acid to another" (Koonin, Novozhilov 2017, 48).

“Knowledge transfer and convergence of methodological practices: Cases of interdisciplinary integration of political, biological, and linguistic research” at the Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences.

## REFERENCES

- Barbieri, M., 2012. Codepoiesis—the deep logic of life. *Biosemiotics* 5(3), 297–299.
- Barbieri, M., 2015. *Code Biology. A New Science of Life*. Dordrecht: Springer.
- Barbieri, M., 2018. What is code biology? *BioSystems* 164, 1–10.
- Barbieri, M., 2019. A general model on the origin of biological codes. *BioSystems* 181, 11–19.
- Barbieri, M., 2019. Evolution of the genetic code: The ambiguity-reduction theory. *Biosystems* 185, available at: <<https://doi.org/10.1016/j.biosystems.2019.104024>>.
- Chebanov, S. V., 2019. Steps towards the semiotic awareness of biology: Biosemiotics replacing the role of synthetic theory of evolution. *METHOD: Moscow Yearbook of Works from Social Sciences / Russian Academy of Sciences, INION* 9. 151–172 (In Russian).
- Copley et al., 2005. A mechanism for the association of aminoacids with their codons and the origin of the genetic code. *Proc. Natl. Acad. Sci* 102, 4442–4447.
- Crick, F. H., 1968. The origin of the genetic code. *J Mol Biol.* 38(3), 367–379. doi:10.1016/0022-2836(68)90392-6.
- Crick, F., 1981. *Life Itself: Its Origin and Nature*. NY: Simon & Schuster.
- Crick, F. H., Griffith, J. S., Orgel. L. E., 1957. Codes without commas. *Proc. Natl. Acad. Sci* 43(5), 416–421.
- Frank-Kamenetsky, M. D., 1980. Genetic codes. *Khimiya i Zhizn (Chemistry and Life)* 5, 34–41 (In Russian).
- Frank, A., Froese, T., 2018. The Standard Genetic Code can Evolve from a Two-Letter GC Code Without Information Loss or Costly Reassignments. *Origins of Life and Evolution of Biospheres* 48, 259–272. doi: 10.1007/s11084-018-9559-4.
- Gimona, M., 2008. Protein linguistics and the modular code of the cytoskeleton. In Barbieri, M., Hoffmeyer, J. (Eds.), *Codes of life: The rules of macroevolution*. Berlin: Springer, pp. 189–206.
- Gusev, V. A., Schulze-Makuch, D., 2004. *Physics of Life Reviews* 1. pp. 202–229.
- Higgs, P. G., 2009. A four-column theory for the origin of the genetic code: Tracing the evolutionary pathways that gave rise to an optimized code. *Biology Direct* 4(1), available at: <<https://doi.org/10.1186/1745-6150-4-16>>.
- Hoffmeyer, J., 2007. Semiotic scaffolding of living systems. In Barbieri, M. (Ed.), *Introduction to Biosemiotics: The New Biological Synthesis*. Dordrecht: Springer, pp. 149–166.
- Ikehara, K., 2002. Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. *J. Biosci.* 27, 165–186.
- Jakobson, R., 1970. Linguistics. In Unesco (Ed.), *Main trends of research in the social and human sciences, Part one: Social sciences*. The Hague, Paris: Mouton, pp. 419–463.
- Jiménez-Sánchez, A., 1995. On the origin and evolution of the genetic code. *J Mol Evol* 41, 712–716.
- Koonin, E. V., Novozhilov, A. S., 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61(2), pp. 99–111. doi: 10.1002/iub.146.
- Koonin, E. V., Novozhilov, A. S., 2017. Origin and Evolution of the Universal Genetic Code. *Annual Review of Genetics* 51, 45–62.
- Kun, A., Radvanyi, A., 2018. The evolution of the genetic code: impasses and challenges. *Biosystems* 164, 217–225, available at: <[doi.org/10.1016/j.biosystems.2017.10.006](https://doi.org/10.1016/j.biosystems.2017.10.006)>.
- Kull, K., 2015. Evolution, choice, and scaffolding: Semiosis changing its own building. *Biosemiotics* 8, 223–234, available at: <<https://doi.org/10.1007/s12304-015-9243-2>>.
- Kull, K., 2020. Codes: Necessary, but not sufficient for meaning-making. *Constructivist Foundations* 15(2): 137–139.
- Lacková, L., 2018. A Biosemiotic Encyclopedia: an Encyclopedic Model for Evolution. *Biosemiotics* 11, 307–322, available at: <<https://doi.org/10.1007/s12304-018-9325-z>>.
- Lacková, L., Matlach, V., Faltýnek, D., 2017. Arbitrariness is not enough: towards a functional approach to the genetic code. *Theory Biosci* 136, 187–191, available at: <<https://doi.org/10.1007/s12064-017-0246-1>>.
- Lane, N., 2009. *Life Ascending: The Ten Great Inventions of Evolution*. NY: Norton & Co.
- Monod, J., 1971. *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. Fontana.
- Patel, A., 2005. The triplet genetic code had a doublet predecessor. *Journal of Theoretical Biology* 233(4), 527–532.
- Pattee, H., 2005. The necessity of biosemiotics: Matter-symbol complementarity. *Journal of Biosemiotics* 1(1), 223–238.
- Penny, D., 2005. An Interpretive Review of the Origin of Life Research. *Biol Philos* 20, 633–671, available at: <<https://doi.org/10.1007/s10539-004-7342-6>>.
- Ratner, V. A., 2000. The Chronicle of the great discovery: ideas and persons. *Priroda*. 6, 22–30 (in Russian).
- Rodin, A. S., Szathmáry, E., 2011. On origin of genetic code and tRNA before translation. *Biol Direct*. 6, 14. doi: 10.1186/1745-6150-6-14.
- Rumer, Y. B., 1966. About the codon’s systematization in the genetic code. *Proc. USSR Acad. Sci.* 167, 1393–1394 (in Russian).
- Saussure, F. de, 1959[1916]. *Course in General Linguistics*. Translated, with an introduction and notes by Wade Baskin. New York, Toronto, London: McGraw-Hill Book Company.

## Zolyan

- Sengupta, S., Higgs P. G., 2015. Pathways of genetic code evolution in ancient and modern organisms. *J. Mol. Evol.* 80, 229–243. doi: 10.1007/s00239-015-9686-8.
- Štambuk, N., Konjevoda, P., Turčić, P., et al., 2018. Genetic coding algorithm for sense and antisense peptide interactions. *Biosystems* 164, 199-216. doi: 10.1016/j.biosystems.2017.10.009.
- Taylor, F. J., Coates, D., 1989. The code within the codons. *Biosystems* 22(3), 177–187. doi: 10.1016/0303-2647(89)90059-2.
- Trifonov, E., 2004. The Triplet Code From First Principles. *Journal of Biomolecular Structure and Dynamics* 22(1), 1–11.
- Trifonov, E. N., 2009. The origin of the genetic code and of the earliest oligopeptides. *Research in Microbiology* 160, 481–486.
- Wu, Huan-Lin, Bagby, S., Elsen J. M. H., 2005. Evolution of the Genetic Triplet Code via Two Types of Doublet Codons. *J Mol Evol* 61(1), 54–64. doi: 10.1007/s00239-004-0224-3.
- Zolyan, S., 2018. The Genetic code: Grammar, semantics, evolution. *METHOD: Moscow Yearbook of Works from Social Sciences / Russian Academy of Sciences, INION* 8. 130–184 (In Russian).
- Zolyan, S., Zdanov, R., 2018. Genome as (hyper)text: From metaphor to theory. *Semiotica* 2018(225), 1–18, available at: <<https://doi.org/10.1515/sem-2016-0214>>.

[Return to front page ↑](#)