2
Are the Strange Information Structures of the Genetic Code an Accident or an Artifact?

Alexander D. Panov

Skobeltsyn Institute of Nuclear Physics of Lomonosov Moscow State University, Moscow, Russia

Felix P. Filatov

I. I. Mechnikov Research Institute of Vaccines and Sera; Gamaleya National Research Center for Epidemiology and Microbiology, Moscow, Russia

Abstract

The universal genetic code has several strange symmetries, the biochemical significance of which remains unclear. In addition, it is possible to construct many numerical signatures, usually related to divisibility by 111, from the weights of the amino acids encoded by the genetic code. This article is based on the work of Vladimir Shcherbak and Maxim Makukov (2013), in which many such independent information structures were obtained for the first time. In the current work several new information structures are found, and it is shown that their entire collection naturally divides into five 'information levels', the first two of which are particularly simple and resemble the 'attention signals' known in the practice of searching for signals of extraterrestrial intelligence in the SETI problem. The method of analysis used in the article differs significantly from the Shcherbak – Makukov approach and is based on the systematic use of the 'simplicity criterion'. The article presents in detail both new and known information signatures of the genetic code and discusses the prospects for resolving the question of the accident or artificial nature of these strange information structures.

Keywords: *origin of life, astrobiology, exobiology, genetic code.*

Evolution: Environmental, Demographic, and Political Risks 2024 23–68
DOI: 10.30884/978-5-7057-6399-3_03

1. Introduction

The title of the article should be taken literally: the content of the article is a detailed explanation of the question posed in the title. No definite answer to this question is given, but possible scenarios of the development of the topic are considered. The paper establishes the existence of the alternative 'accidental/artificial' with respect to strange information structures that do exist in the genetic code and argues the validity of the question posed in the title of the article. Artificiality of the mentioned information structures (the second of the possibilities presented in the alternative) means extraterrestrial origin of terrestrial life, so it brings the whole problem into the field of astrobiology and SETI problems.

The idea that the genetic code may contain information signaling about artificial interference in the origin of terrestrial life was first formulated by G. Marx (1979). He wrote that the genetic code, because of its extreme stability and conservatism, is the most reliable carrier of information of a biological nature, although the amount of information it can contain is small (shorter than 40-letters long text in a conventional alphabet). Marx noted, however, that at the time of his writing no evidence of information other than coding rules had been found in this carrier. Serious searches for such information in the structures of the genetic code were undertaken in the works of Vladimir Shcherbak and Maxim Makukov (Shcherbak and Makukov 2013; Makukov and Shcherbak 2018), and these searches led to the discovery of a large number of non-trivial information patterns in it. These articles served as the starting point for the present work.

What could the signature of artificiality in a genetic code look like? It could be numerical relationships associated to the structure of the code, that do not seem accidental, but whose natural origin has no explanation; unexpected symmetries of the code, *etc.* The search for such 'peculiarities' of the code may cause a reproach to numerology. However, what else could such information signatures look like? Here it is important not to fall into unwitting fitting and manipulation of numbers and to use the relations 'lying on the surface', to avoid too complex and artificial combinatorial constructions and to pay attention to what in a certain sense satisfies the criterion of 'naturalness and simplicity'. At least this is where we should start.

The articles by Shcherbak and Makukov, however, start immediately with rather complicated constructions. For our part, we tried to follow strictly the aforementioned criterion of naturalness and simplicity explicitly. Our approach also allowed us to notice a new feature of unusual information patterns found in the genetic code: they are naturally arranged on several information levels in the order of increasing complexity and in the order of attracting new resources for information representation. Moreover, the first two simplest levels resemble the

'attention signal' that is often assumed to be present in messages from extraterrestrial intelligence, and which is actually present in real terrestrial messages already sent to extraterrestrial intelligence from us (Zaitsev 2008). Our approach also provided additional support for a more sophisticated approach to the topic, which was used in the work of Shcherbak and Makukov.

In this paper, we will consistently present all known to us non-trivial information patterns of the genetic code. For the sake of brevity, we will simply call such structures signatures, without implying that they necessarily mean anything. Not all the constructions in Shcherbak and Makukov's work seem to us to be flawless. In our review we will present some new signatures as well as signatures found by Shcherbak and Makukov themselves (partly reinterpreted), which are not objectionable, as well as signatures known from even earlier works (related to the so-called Rumer symmetry, see below). In the course of the review we will clarify whether the signature under discussion is new or already known from previous studies. The main order of the presentation will follow the increasing complexity of the information levels mentioned above.

Our presentation will conclude with a discussion of the results obtained and further perspectives.

We will begin with a short presentation of the structure of the genetic code, which will also allow us to introduce the necessary terminology.

2. The Genetic Code Table

The table of the genetic code, known from all textbooks (see Fig. 1), was proposed by Francis Crick (1968). Since each coding product (amino acid or **stop** signal) is defined by a triplet of nucleotides (bases), also called a codon, the code table would have the form of a three-dimensional matrix of $4 \times 4 \times 4$, indexed on each side by four nucleotides, for example, in the order Thymine-Cytosine-Adenine-Guanine (T-C-A-G in single-letter notations, see Fig. 2). Each triplet of nucleotides corresponds to one cell in the matrix, where the amino acid encoded by that codon or the stop signal is located. An ordinary two-dimensional table or matrix has rows and columns located in the horizontal plane of a 'sheet of paper', in a three-dimensional matrix vertical columns are added (see Fig. 2). We can say that a three-dimensional matrix is a two-dimensional matrix with vertical columns in its cells. A three-dimensional matrix is inconvenient to represent on a two-dimensional sheet of paper. Thus, in Fig. 1 the third dimension corresponding to the third base in the codon is not placed as a vertical column under a cell of the two-dimensional matrix, as in Fig. 2, but 'lying down', so that each of the cells of the two-dimensional table indexed by the first two bases of the codon is itself a column of four cells that are numbered by the third base. Since the contents of these larger cells actually come from the vertical columns of the three-dimensional matrix, we will still

refer to them as columns. Thus, each column contains four coding products at constant first two codon bases and a changing third base.

		SECOND BASE			
		T	C	A	G
FIRST BASE	T	TTT F TTC	TCT S TCC	TAT Y TAC	TGT C TGC
		TTA L TTG	TCA TCG	TAA **stop** TAG	TGA **stop** TGG W
	C	CTT L CTC CTA CTG	CCT P CCC CCA CCG	CAT H CAC CAA Q CAG	CGT R CGC CGA CGG
	A	ATT I ATC ATA	ACT T ACC ACA	AAT N AAC AAA K	AGT S AGC AGA R
		ATG M	ACG	AAG	AGG
	G	GTT V GTC GTA GTG	GCT A GCC GCA GCG	GAT D GAC GAA E GAG	GGT G GGC GGA GGG

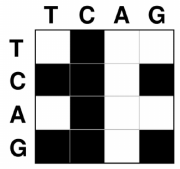


Fig. 1. Table of the universal genetic code in Crick's form and its calligram. The coding products (amino acids) are indicated by one-letter symbols, the triplets terminating translation are indicated by the word **stop**. The gray cells refer to Rumer octet **I**, the rest refer to octet **II** (see text)

The bases T, C, A, G are represented by compounds of two types: T and C are pyrimidines, A and G are purines. We will use the standard notation Y for pyrimidines and R for purines. Furthermore, we will denote the set of bases (T, C, A) by H, and the whole set of bases (T, C, A, G) by V (see Table 1).

In the table of the genetic code in Fig. 1, the standard one-letter notations **F**, **L**, ... are used to label amino acids. Table 2 shows the full names of all 20 amino acids in the genetic code together with their standard one-letter and three-letter notations. The table also gives the total atomic weight *M* of each free amino acid (otherwise called mass number): the sum of atomic weights of its constituent atoms (for each atom, the number of protons together with neutrons in the nucleus), and for each atom the atomic weight of its most abundant stable isotope is taken: for hydrogen 1, for carbon 12, for nitrogen 14, for oxygen 16, and so on.

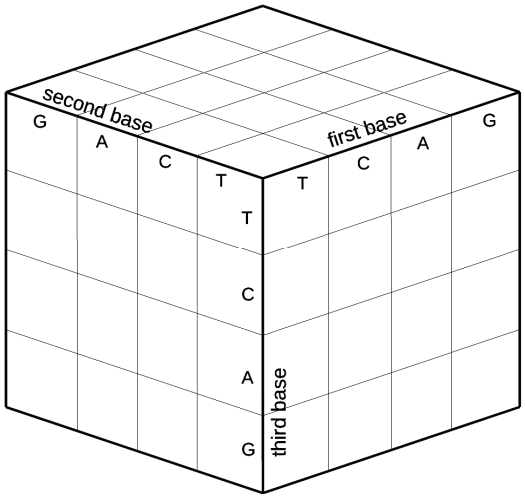


Fig. 2. Three-dimensional matrix indexed by bases T, C, A, G

Each amino acid consists of a constant part and a side chain (see Fig. 3). The constant parts of all amino acids with the sole exception of proline (**P**, **Pro**) are the same and have an atomic weight of 74. The constant part of proline has an atomic weight of 73. The side chains have a variety of atomic weights M_{Side}, which are also listed in Table 2.

Table 1. Notations for groups of bases

Notation	Group of bases
Y	(T, C) – pyrimidines
R	(A, G) – purines
H	(T, C, A)
N	(T, C, A, G) – any base
M	(T, G) – the first Rumer pair
K	(C, A) – the second Rumer pair

As can be seen from the table of the genetic code in Fig. 1, in most cases several different codons correspond to one coding product. This is the so-called degeneracy of the genetic code. In some cases, all codons of one column encode the same product. We will call such columns homogeneous. In Fig. 1, they are marked in gray. In other cases one column encodes different products, we

will call them heterogeneous, such columns are left uncolored. We will compare a simple picture with the table of genetic code, which we will call a calligram. A calligram is a 4 × 4 square, the cells of which are shaded black if they correspond to homogeneous columns of the genetic code table (gray in Fig. 1), and left unshaded if they correspond to heterogeneous columns (unshaded in Fig. 1). The calligram corresponding to the basic genetic code table is shown in Fig. 1 on the right. Later we will consider other representations of the genetic code table, to which other calligrams will correspond. Let us now proceed step by step to the representation of the non-trivial information structures of the genetic code.

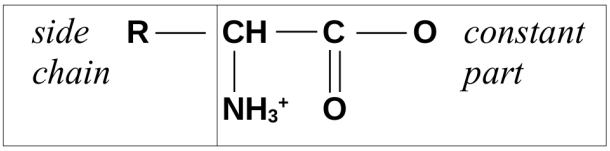


Fig. 3. Amino acid structure: constant part and side chain

Table 2. Amino acids and their atomic weights

Amino acid			M_{Side}	M	Amino acid			M_{Side}	M
Glycine	**G**	**Gly**	1	75	Asparginic acid	**D**	**Asp**	59	133
Alanine	**A**	**Ala**	15	89	Glutamine	**Q**	**Gln**	72	146
Serine	**S**	**Ser**	31	105	Lysine	**K**	**Lys**	72	146
Proline	**P**	**Pro**	42	115	Glutamic acid	**E**	**Glu**	73	147
Valine	**V**	**Val**	43	117	Methionine	**M**	**Met**	75	149
Threonine	**T**	**Tre**	45	119	Histidine	**H**	**His**	81	155
Cysteine	**C**	**Cys**	47	121	Phenylalanine	**F**	**Phe**	91	165
Leucine	**L**	**Leu**	57	131	Arginine	**R**	**Arg**	100	174
Isoleucine	**I**	**Ile**	57	131	Tyrosine	**Y**	**Tyr**	107	181
Aspargin	**N**	**Asn**	58	132	Tryptophan	**W**	**Trp**	130	204

3. The First Information Level. Rumer Symmetry and Related Symmetries of the Genetic Code

Without going into speculation about the origin of the code, Francis Crick called it a ‘frozen accident’. Later, three different theories were proposed that explain the structure of the code to a certain extent (Nikitin 2016): the theory of optimization for the minimum of protein synthesis errors, the theory of struc-

tural correspondence of amino acids to codons (key-lock), and the theory of coevolution of codons and amino acid biosynthesis pathways. However, it turned out that the genetic code has some unexpected formal properties, that, to the extent of common understanding, cannot be explained in a simple and unambiguous way within the framework of modern concepts related to biochemistry or evolution.

Yuri Rumer was the first to draw attention to them (Rumer 2013; Konopelchenko and Rumer 1975). Trying to find a rational basis for the structure of the genetic code, he combined homogeneous columns into one set and heterogeneous columns into another. The number of columns in each set was eight (see Fig. 1), so these two sets were called octets; we will refer to them with the Roman numerals **I** and **II**. Octet **I** corresponds to homogeneous columns, octet **II** to heterogeneous columns. Rumer's classification is not trivial, since the ratio of the number of columns in the two sets could well be different, and indeed is different in some existing rare alternative genetic codes (see Section 9.3.).

Quite unexpectedly, octets **I** and **II** turn out to be related by a simple symmetry transformation: R = (T ↔ G, C ↔ A), which we will call the Rumer transformation. In this base swapping, each column of octet **I** goes to some column of octet **II** and vice versa. For example, the CT column of octet **I** goes to the AG column of octet **II**, the AG column goes to CT, and so on.

The presence of this strange symmetry is, on the one hand, easy to detect, as well as the very existence of the octets, but on the other hand, it cannot be understood from the point of view of the special role of this symmetry in the functioning of the genetic code. If we assume that in the presence of two octets, the rest of the genetic code is randomly arranged, then the conditional probability of the appearance of the Rumer symmetry is 1/256, *i.e.* this symmetry looks rather unlikely, fragile and optional in a functional sense (which is confirmed by its absence in some alternative genetic codes).

It should be noted that this symmetry has been independently rediscovered at least twice since Rumer (Danckwerts and Neubert 1975; Wilhelm and Nikolajewa 2004), because it is easy to detect, as it literally ‘lies on the surface’, but it does not attract the attention of biochemists, as the link between this symmetry and the functions of the genetic code is not recognized.

The Rumer symmetry is now well known. It is discussed in particular in the works of Shcherbak and Makukov. Let us take a new step in the analysis of these symmetries. Note that Fig. 1 shows only one of the possible ways of representing the genetic code table. Other natural ways of representation can be obtained by changing the order of the bases (T-C-A-G) on the sides of the matrix to another order, *i.e.,* by subjecting this row to a permutation.

As it is easy to understand, when rearranging the bases on the sides of the matrix, the columns of the table will somehow change places, and within the columns the third base of the codons will also appear in a different order, but the full set of coding products in the columns will remain unchanged. Therefore, if some column belonged to the octet **I**, it will still belong to the octet **I** after any rearrangement of the bases on the sides of the table, and the columns of the octet **II** will not change their nature either. Therefore, the Rumer octets are invariant to transformations given by permutations in a string of four symbols: under such transformations, the octets **I** and **II** pass into themselves. Any two successive permutations give some new permutation, there is an identity permutation which leaves the base string unchanged, and every permutation has a reverse permutation which returns the base string to its original state. All this means that the permutations form a group of transformations,[1] and the group of permutations of four elements is a symmetry group of Rumer octets in the sense that Rumer octets do not change under transformations of this group.

Fig. 1 shows that the structure of the regions of the calligram corresponding to octets **I** and **II** is rather complicated. The question arises: is it possible to visualize it by rearranging the order of the bases on the sides of the table, so that the regions corresponding to octets **I** and **II** take a simpler form? In particular, in Fig. 1, each of the regions **I** and **II** is not connected (moreover, each of them consists of three separate parts). Could it be that there are such arrangements of bases on the sides of the table of the genetic code that each of the regions corresponding to octets **I** and **II** were connected, *i.e.* had no breaks?

To answer these questions, it is necessary to analyze the calligrams corresponding to all the possible arrangements of bases on the sides of the table of the genetic code. There are in total $4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$ such variants. It is not necessary to consider all 24 calligrams because half of them can be obtained by inverting the calligram with respect to the center of the square (matrix), which is also equivalent to rotating the matrix by 180º. Such an operation does not change the figure in essence.

---

[1] A group is a mathematical structure, denoted G, which is a set over whose elements a binary operation, usually called multiplication, is defined and which has the following properties: 1) there exists a unit element $\boldsymbol{e}$ such that for any element $\boldsymbol{a}$ of G $\boldsymbol{a} \cdot \boldsymbol{e} = \boldsymbol{e} \cdot \boldsymbol{a} = \boldsymbol{a}$; 2) for any element $\boldsymbol{a}$ of G, there exists an inverse element $\boldsymbol{a}^{-1}$ such that $\boldsymbol{a} \cdot \boldsymbol{a}^{-1} = \boldsymbol{a}^{-1} \cdot \boldsymbol{a} = \boldsymbol{e}$; 3) multiplication is associative: $\boldsymbol{a} \cdot (\boldsymbol{b} \cdot \boldsymbol{c}) = (\boldsymbol{a} \cdot \boldsymbol{b}) \cdot \boldsymbol{c}$. The main applications of group theory are related to the fact that a variety of transformations, such as rotations of a figure, translating a figure into itself (called the symmetry of that figure), permutations of symbols in a string, *etc.* form groups. A group multiplication is a consecutive application of two transformations, the unit of a group is an identical transformation that changes nothing, and an inverse transformation for any transformation is one that returns the object to its original form, *e.g.,* for a rotation of a figure – a rotation by the same angle in the opposite direction, *etc.*

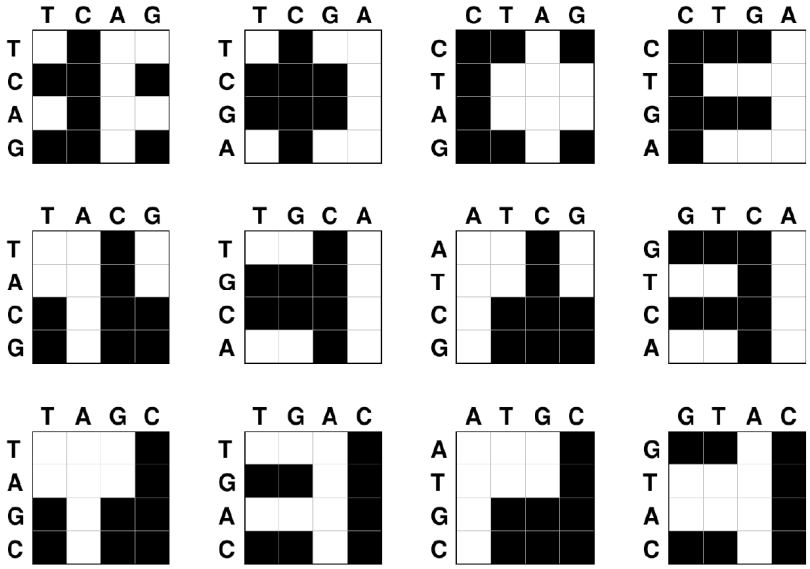


Fig. 4. All 12 non-trivial calligrams corresponding to all possible representations of the table of the genetic code. Another 12 calligrams are obtained by rotating the calligrams in the figure by 180º (or by inversion relative to the center), which corresponds to the mirror reflection of the sequence of bases of each calligram

Fig. 4 shows all 12 non-trivial calligrams, from which the other 12 can be obtained by 180º rotation. The calligram corresponding to the original representation of the table in Fig. 1 is in the upper left corner of Fig. 4. It can be seen that this calligram is not very simple compared to the others. Only two calligrams are particularly simple: CTGA and ATGC. Their simplicity is expressed in the fact that each of the octets **I** and **II** in these calligrams is represented by a connected area, that is a continuous area without breaks.

It turns out that all the connected calligrams (there are four of them: two, which are shown in Fig. 4, and two more, which are obtained from them by rotating by 180º) are related by the permutation of bases corresponding to the Rumer transformation and through two more transformations directly related to the Rumer transformation. Along with the Rumer transformation R = (T ↔ G, C ↔ A), let us consider two more transformations that are in some sense 'halves' of the Rumer transformation: R_1 = (T ↔ G), R_2 = (C ↔ A). In the group-theoretic sense, $R = R_1 \cdot R_2$, $R_1 = R \cdot R_2$, $R_2 = R \cdot R_1$, and it is easy to

check that all three transformations together with the identity transformation form a group (it is a subgroup of the group of all permutations) which turns out to be the symmetry group of connected calligrams: the property of connectedness is an invariant of this group of transformations. Fig. 5 shows how all connected calligrams are transformed through each other by the transformations R, R_1, R_2 (in mathematics, figures of this kind are called commutative diagrams). Not only do all connected calligrams turn out to be related to each other by means of a simple group of transformations directly related to the Rumer transformation, but the picture itself (see Fig. 5) has an amazing symmetry. It is easy to see that it is symmetric with respect to inversion or with respect to 180^o rotation. The calculations show that the probability of such additional symmetry in the presence of Rumer symmetry is 1/4, that is the total probability of obtaining Rumer symmetry together with the symmetry of the connected calligrams, in the presence of Rumer octets, is $1/256 \times 1/4 = 1/1024$.

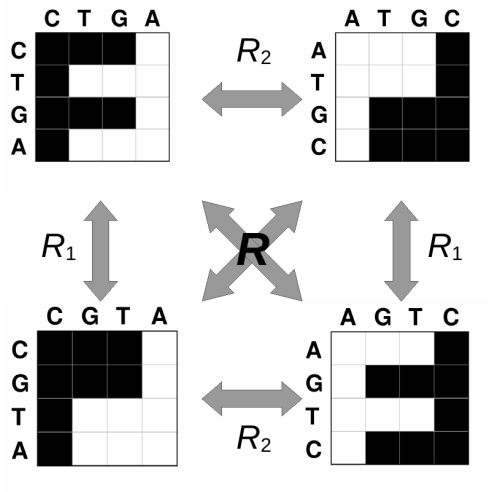


Fig. 5. Action of the group of transformations of connected calligrams (see text)

As long as we could not see any hidden formal meaning in the Rumer symmetries, all this could be interpreted as an amusing accident. However, these surprising symmetries can serve as a kind of 'attention signal': they suggest that octets **I** and **II** should be examined more closely. Note that at this stage it is not even too important how 'strange' the Rumer symmetries are in the sense of small probabilities favoring their occurrence; what is important is that the presence of these symmetries inevitably draws attention to Rumer octets as such.

4. Second Information Level. Signatures of Full Weights of Coding Products

Taking into account the 'hint' received from the side of the Rumer symmetry, let us deepen the analysis by attracting new types of data. From this point of view, all the information associated with the Rumer symmetries and with additional symmetries of connected calligrams (see Section 3) represent the first level of information signatures of the genetic code. Our advance into the depth is justified by the above-mentioned 'attention signal', but we will still avoid complex constructions and consider only signatures of maximum 'simplicity and naturalness'.

For this purpose, we turn our attention to the full masses of the coding products (see Table 2). In order to count the masses of any group of encoding products, we would like to assign a weight to each encoding product, but a non-trivial problem immediately arises here. The **stop** signal is also a coding product, but it is not a material object, it is a purely logical notion, it cannot naturally be assigned any weight, not even weight zero. Some things actually do have weight zero, for example, a photon, but the logical abstraction **stop** has no weight, which is simply not one of the properties of the concept **stop**. Therefore, we must artificially attach some weight to the **stop** signal. A fundamental and non-trivial ambiguity arises here (Shcherbak and Makukov did not pay attention to it). Let us emphasize this circumstance, since we will have to make use of it later. It seems that the simplest and most natural solution is to assign weight 0 to the **stop** signal, which we will do for now. Here it is important to realize that we have made a fundamentally important logical jump. The sums of the masses of the coding products, which we will deal with below, are now not real physical quantities. If the calculated sum of masses includes the 'weight' of the **stop** signal, then this sum is a purely logical concept, and there is no point in even asking what biochemistry can explain such a sum. On a slightly different topic, similar explanations are given in the work by Shcherbak and Makukov (2013) (see Section 5).

From our point of view, such a strong logical step would be completely inadmissible if the analysis were to start immediately with the full weights of the coding products without having the 'attention signal' from the Rumer symmetries side. It is the presence of such a 'signal' that justifies such a non-trivial logical step. The presence of this special logical step allows us to say that we have moved to a new level of complexity of analysis, which has been called the second information level in the title of this section. Another sign of the new information level is the inclusion of a new resource – the full weights of the coding products.

Now we notice that all coded amino acids consist of the so-called constant part and the side chain (see Fig. 3). The constant part of all amino acids, with a single exception, is the same and has a weight of 74, but the radicals are all different. This single exception is the amino acid proline **P** (weight of the constant part is 73), but the special role of the number 74 in our whole problem is

quite obvious, because 19 out of 20 amino acids have exactly this value of the weight of the constant part.

Now let us do a very simple procedure. Let us calculate the total weight of the products coded in each octet separately. In this case, we will count each product exactly as many times as it appears in the table (*i.e.,* we count each cell separately). This seems to be the simplest and most natural way to count. For example, for the first octet column CT, the product **L** will be counted four times, for the second octet column TG, the product **C** will be counted twice, the **stop** signal will be counted once and **W** will be counted once, and so on. This gives a total weight number of 3700 for the first octet and 4218 for the second octet. By their very nature, these numbers are *a priori* random integers, since they are obtained by summing up several integers, the masses of different amino acids, which are unrelated to each other. The masses of the amino acids do not obey any simple regularity (see Table 2).

It turns out that each of these two numbers 3700 and 4218 is divisible by the 'magic number' 74, which is already fixed *a priori* by the context of the problem. There is no reason to expect that the full masses of the coding products of the octets constructed above will be integer multiples of the weight of the constant part assigned *a priori* in the context of this problem, which is also, in a certain sense, a random number. Given the random nature of the numbers 3700 and 4218, we can estimate the probability of this strange coincidence. The probability that a random integer is divisible by 74 is 1/74, while the probability that two independent random integers are both divisible by 74 is only 1/74 × 1/74 = 1/5476. All of this looks strange, to say the least, although a chance cannot be ruled out, of course. This signature was not noticed in the work of Shcherbak and Makukov, since they did not work with full masses of coding products. Their approach was more sophisticated from the very beginning (see Section 5).

But that is not all. The weight of the second octet turns out to be of the special form 4218 = 4 × 999 + 222. The numbers of this form we will call Shcherbak-Makukov numbers (see Section 5). They play a leading role in the information signatures of the following levels.

But that is not all again. The first Rumer octet is simpler than the second octet and seems to be in some sense more fundamental. If we arrange the coding products of the first octet in order of increasing weight (what could be simpler?), the first bases of the corresponding codons form the sequence GGTC| GACC, which is mirror-symmetric with respect to its center (marked as |) in terms of complementarity;[2] nucleotide C is complementary to nucleotide G, nucleotide T is complementary to nucleotide A, and so on. The probability cor-

[2] Recall that in the formation of the DNA double helix, the structure of two adjacent strands is linked by the complementarity relation of the bases: G↔C, T↔A. That is, opposite to **G** there is always **C**, opposite to **T** there is always **A**.

responding to this 'strangeness' is not quite easy to estimate. What arises here is a special case of the so-called 'corner search problem' (for more details see Section 9.2.). Applied to this situation, the problem looks like this. It is not difficult to calculate the probability of occurrence of mirror-complementary symmetry under the assumption of complete randomness of the sequence of bases, it is 1/256. But we should take into account that besides the existing symmetry there are other similar symmetries that would surprise us no less. For example, if there were not a mirror symmetry, but an exact repetition of the first four letters by complementarity. These are already two possibilities. Besides, it is possible to include in the analysis other relations between bases besides complementarity, among them at least Rumer's transformation pairs, complete identity of bases and transpositions inside the purine and pyrimidine pairs. Each of these relations is also included twice: for mirror symmetry and for exact repetition. In total, we have already obtained eight similar symmetries, so the probability obtained above should be multiplied by 8, we get 8/256 = 1/32.

Interestingly, if we consider only whether the base in the GGTCGACC string is purine (R) or pyrimidine (Y), then in addition to the already existing complementary mirror symmetry, we obtain a structure with shift symmetry: RRYYRRYY. Both these symmetries have been recognized in the article by Shcherbak and Makukov (2013).

Thus, we descended one information level deeper in the complexity of the analysis and, as in the first information level, we did not come up empty-handed. Let us emphasize once again that in order to obtain numerical and symmetric patterns of the second level of complexity, we did not have to perform any complicated manipulations with digits, except for the artificial assignment of zero weight to the **stop** signal (which seems very simple at first glance). All the results are 'on the surface' and are obtained naturally. In fact, the information patterns of the second level are so simple that we can say that both information levels play the role of 'attention signals' in a sense. However, with the second information level, we have obtained a significant amplification of the 'attention signal', which allows us to start looking for more subtle and complex information patterns of the genetic code. This provides a moral justification for more complex manipulations of the data under discussion.

5. The Third Information Level. Masses of Side Chains

The main information signature array of the article (Shcherbak and Makukov 2013) was not obtained using the full masses of amino acids, as we obtained the masses of octets **I** and **II** (3700 and 4218) at the second information level, but using the masses of amino acid side chains. Here, however, one surprising circumstance is revealed. A significant part of the information patterns appears only if one artificially adjusts the structure of the proline molecule, which has a non-standard weight of the constant part: 73 instead of 74. It is

necessary to artificially (virtually!) transfer one hydrogen from the side chain of proline to the constant part, after which the weight of the side chain becomes 41 instead of the original 42, and the weight of the constant part becomes standard, that is 74. Shcherbak and Makukov call this mental operation an 'activation key', which gives access to the main array of signatures of the genetic code. As correctly noted in the work by Shcherbak and Makukov (2013), all signatures acquire a virtual character: various sums of amino acid side chain weights, taking into account the proline correction are not something real, so the question of why these sums are such and not others loses its physical meaning.

A problem similar to the one already discussed above arises again with the **stop** coding product. If we now consider not the full masses of the amino acids, but only the masses of the side chains, what weight of 'side chain' should be assigned to the **stop** signal? Shcherbak and Makukov artificially assign a weight zero to the **stop** signal, and it is this and only this choice that leads to the large set of information patterns in the article (*Ibid.*). In essence, not one 'activation key' is used, as Shcherbak and Makukov believe, but two. The transition to the use of amino acid side chain weights together with the now two artificial adjustments marks the transition to the third information level of complexity.

The artificial assignment of zero weight to the 'side chain' of the **stop** signal has some interesting consequences that were not considered by Shcherbak and Makukov. After standardizing of the weight of the constant part, all masses of amino acid side chains are calculated as the total weight of the molecule minus 74. For uniform behavior of all coding products, one would assume that the same should be true for the **stop** signal. But in order to get zero for the **stop** side chain weight, we have to assume that the 'total weight' of **stop** is not zero at all, as we assumed at the second information level (see Section 4), but 74. Only in this case we obtain the correct value for the 'side chain' of the **stop** signal: $74 - 74 = 0$. The possibility of redefinition of the weight of **stop** signal is consistent with the freedom to choose the full weight of the product **stop**, as we wrote above (see Section 4). Although the choice of zero value for the total weight of the **stop** signal seems to be the simplest, it is in a sense not the most logical (or correct).

With this new understanding of the nature of the 'total weight' of the **stop** signal, we can return to the calculation of the total masses of the Rumer octets (see Section 4). Nothing changes for the first octet with its total weight of 3700, since the **stop** signal is not included in the calculation of its weight, but the weight of the second octet changes from 4218 to 4440. This is the place where a small probabilistic miracle occurs. While 3700 and 4218 had a greatest common divisor of 74, which was already an unexpectedly large number, 3700 and 4440 have a greatest common divisor of 740! It turns out that assigning a total

weight of 74 to the **stop** signal not only corresponds to the 'activation key' of the signatures discovered by Shcherbak and Makukov, but also significantly strengthens the signature of the second information level, one of the two 'attention levels'. But this, it turns out, is not all.

Let us decompose the new common divisor 740 into multiples as follows:

$$740 = 2 \times 10 \times 37. \quad (1)$$

Now let us notice that in our problem there are two numbers highlighted from the very beginning – the weight of the constant part of amino acids 74 and the number of amino acids appearing in the genetic code 20. Let us write these numbers as follows:

$$20 = 2 \times 10;\ 74 = 2 \times 37. \quad (2)$$

In all the numbers on the right-hand sides of the equations (1, 2) there is a common multiplier 2, by which we reduce these numbers, after which we get the following series of numbers: 10 × 37, 10, 37. The pair of numbers (10, 37) appears in two independent ways at once: the first time as co-multipliers of the number 370, the second time separately. This may indicate some special role of this pair of numbers.[3]

One could say that all this is no more than an exercise in numerology, if this pair of numbers did not play an absolutely exclusive role in the generation of all the basic set of signatures of the genetic code found by Shcherbak and Makukov. Moreover, consideration of the special properties of the pair (10, 37) is a necessary prologue to the review of these signatures.

We should start with the fact that the pair of numbers (10, 37) has some remarkable mathematical properties[4] which were discovered by Pacioli in 1508 and have been known for a long time.

Let us write in the table the results of multiplying the number 37 by the natural numbers from 1 to 54 (see Table 3). If we look at three-digit multiplication results, one can see that among them there are all numbers of the form $n \times 111$, where n varies from 1 to 9 (highlighted in bold). Moreover, for all of these numbers the sum of the digits of the result is equal to the multiplier of the number 37 with which this result is obtained: 3 + 3 + 3 = 9, 6 + 6 + 6 = 18 and so on. For all other three-digit multiplication results, the result of each column is obtained from the result of the top row by cyclic permutation of the digits, for example 2 × 37 = 074, 11 × 37 = 407, 20 × 37 = 740. Further, for multipliers from 28 to 54 the pattern is repeated completely, but the summand 1 × 999 is added to all multiplication results. In the next cycle, the addend 2 × 999 ap-

[3] Note that 3700 = 37 × 10 × 10, which also hints at the special role of the numbers 37 and 10, but 10 occurs twice in this decomposition, so we do not consider this 'signature' as a direct indication of the pair (10, 37).

[4] Here our presentation follows the paper (Shcherbak and Makukov 2013) with minimal variations.

pears, and so on. Moreover, in all results of the form $m \times 999 + n \times 111$, the multiplier used to obtain the number is still equal to the sum of the digits of the result in a certain sense, namely, it is equal to $m \times (9 + 9 + 9) + n \times (1 + 1 + 1)$. Numbers of the form $m \times 999 + n \times 111$ play a crucial role in the information signatures of the genetic code. This is the discovery of Shcherbak and Makukov, so we will call such numbers as Shcherbak–Makukov numbers, or SM-numbers for short. Thus, the simplest SM-numbers are of the form $n \times 111$ for $n = 1, \ldots, 9$ followed by numbers of the form $1 \times 999 + n \times 111$ and so on.

Table 3. Symmetry properties of the multiplication table of the number 37 in the base 10 positional number system

1 037	2 074	3 **111**	4 148	5 185	6 **222**	7 259	8 296	9 **333**
10 370	11 407	12 **444**	13 481	14 518	15 **555**	16 592	17 629	18 **666**
19 703	20 740	21 **777**	22 814	23 851	24 **888**	25 925	26 962	27 **999**
28 999+037	29 999+074	30 999+**111**	31 999+148	32 999+185	33 999+**222**	34 999+259	35 999+296	36 999+**333**
37 999+370	38 999+407	39 999+**444**	40 999+481	41 999+518	42 999+**555**	43 999+592	44 999+629	45 999+**666**
46 999+703	47 999+740	48 999+**777**	49 999+814	50 999+851	51 999+**888**	52 999+925	53 999+962	54 999+**999**

Note that at least one SM-number can be found already at the second information level, without using the masses of the side chains of the amino acids. No matter how you calculate the weight of the second Rumer's octet, either with the assumption of the zero weight of the stop signal, or with the assumption of a weight equal to 74, the result turns out to be a Shcherbak–Makukov number:

$$M_{stop} = 0 \Rightarrow 4218 = 4 \times 999 + 222$$
$$M_{stop} = 74 \Rightarrow 4440 = 4 \times 999 + 444.$$

The surprising elegance of the latter variant may be an additional indication of the preference for the choice of 'full weight' 74 for the **stop** signal (these signatures were not noticed by Shcherbak and Makukov).

At first glance, Table 3 presents special properties of the number 37, but in fact this curious arithmetic is only obtained in the positional number system with base 10, that is, these are the properties of the pair of numbers (10, 37). The number pair (10, 37) is not unique in this respect. Any pair of integers of the form $(q, 111_q / 3)$, where $q = 4, 7, 10, 13, \ldots$ and 111_q means 111 in the po-

sitional number system with base q (namely: $111q = 1 \times q^2 + 1 \times q + 1$), has similar properties (in addition to the multiplication table above, they have some other curious properties that we have not touched upon here [see Shcherbak and Makukov 2013]). This means that, although the properties of the pair (10, 37) are interesting, we should not overestimate the degree of unusualness of these properties.

From this point on, we will use a different way of counting weights by codon groups than we used at the second information level (see Section 4), when we counted each cell of the genetic code table separately. By going through the genetic code table in different ways (see below), we will count the weight of each coding product (or the weight of the side chain corresponding to that product) for each column of the table where it occurs only once, without taking into account the degree of degeneracy of that product in that column. This counting rule is rather non-trivial (proposed and used by Shcherbak and Makukov), and there is no *a priori* justification for it. However, the point is that the information signatures given below are obtained in this way and no other, so this counting method is justified *a posteriori*.

Let us now list the information structures of the genetic code obtained using the weights of the side chains of amino acids (the third information level). Most of them have the form of Shcherbak–Makukov numbers or are related to such numbers. There are many such structures.

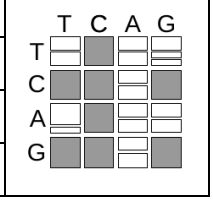
Signature 3.1. Side Chains of the First and Second Octet and the Egyptian Triangle

Let us start with the first Rumer octet, which, as mentioned above, seems to be more fundamental than the second one. The sum of the amino acid side chain weights using the Shcherbak–Makukov counting rule and taking into account the proline 'activation key' (artificial redefinition of the side chain weight and the constant part for proline 42 → 41, 73 → 74) is SM-number 333 (see Table 4):

$$\text{side chains of the octet } \mathbf{I}\text{: } 333 = 37 \times 9. \tag{3}$$

Table 4. The sums of amino acid weights of octet I

amino acid	**G**	**A**	**S**	**P**	**V**	**T**	**L**	**R**	∑
side chain	1	15	31	41	43	45	57	100	**333**
constant part	74	74	74	74	74	74	74	74	**592**
whole weight	75	89	105	115	117	118	131	174	**925**



If we now write out the sum of the weights of the constant parts of the amino acids and the total weights of the amino acids of the octet **I**, we obtain, accordingly:

$$\text{constant parts of the octet } \mathbf{I}\text{: } 592 = 37 \times 16 \quad (4)$$

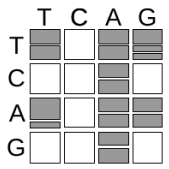
$$\text{full weights of the octet } \mathbf{I}\text{: } 925 = 37 \times 25. \quad (5)$$

All the numbers in equations (3–5) are multiples of 37, and if we reduce them by this common multiplier, we obtain three successive squares of the integers 3^2, 4^2, 5^2, which, as it is hard not to notice, together form the Egyptian triangle: $3^2 + 4^2 = 5^2$. We can also note that this representation of the Pythagorean Theorem does not appear in some accidental place, but specifically in the first octet of Rumer, which resembles the structural center of the genetic code (and is in fact its most stable part), and this signature does not require a complex rule to select the group of amino acids for which the full weight is counted. As will be seen below, the other group selection rules for counting side chain weights, while always making some *a priori* sense, are more complex. It should also be noted that this signature is free from the uncertainty associated with the freedom to choose the weight of the **stop** signal, although it depends critically on the 'activation key' associated with the redefinition of the proline structure. The use of the 'activation key', as already mentioned, is justified *ex post facto* by the fact that it does not lead to a single signature, but to many independent signatures at the same time.

Using the Shcherbak–Makukov counting rule for the sum of the side-chain weights of the coding products of octet **II**, we obtain the sum of the weights, which is also an SM-number: $1110 = 999 \times 1 + 111$ (see Table 5). Recall that we have already obtained the SM-number for octet **II** once: it was the total weight of octet **II** obtained by counting each cell of the genetic code table separately. These two signatures are independent of each other.

Table 5. The sums of amino acid weights of octet II. The order of presentation of the amino acids in the table corresponds to the movement from left to right in each row of the genetic code (see Fig. 1) and from top to bottom in each column of the genetic code

amino acid	**F**	**L**	**Y**	**stop**	**C**	**stop**	**W**	**H**	**Q →**
side chain	91	57	107	0	47	0	130	81	72 →
constant part	74	74	74	0	74	0	74	74	74 →
amino acid	**I**	**M**	**N**	**K**	**S**	**R**	**D**	**E**	∑
side chain	57	75	58	72	31	100	59	73	**999+111**
constant part	74	74	74	74	74	74	74	74	**999+111**



There is probably one more signature associated with octet **II**. If we assume that the total weight of the **stop** signal is zero (thus the weight of the constant part of the **stop** signal is also zero), then the total weight of the constant parts of octet **II** when counting by the Shcherbak–Makukov rule gives the SM-number 999 × 1 + 111. In other words, there is an additional equilibrium of the weights of the side and constant parts of octet **II**: 999 × 1 + 111 = 999 × 1 + 111. For the more 'reasonable' assumption M_{stop} = 74 this equilibrium is broken.

Signature 3.2. Codons with Two Identical Bases

There are 36 codons that contain two identical bases and another base that is different from these two bases. This group of bases can be divided into two halves in such a way that in one group there will be codons in which a pair of identical bases are pyrimidines, and in the other group – purines. Then the sums of the weights of each of the halves are equal (equilibrium of weights) and equal to the Shcherbak–Makukov number 999 (see Tables 6 and 7). This is the first example of selection of amino acid groups for summation of weights, which makes sense a priori, but which is rather complicated in comparison with Signature 3.1.

For completely homogeneous codons (all bases are the same) and completely heterogeneous codons (all bases are different), no similar simple signature is obtained. These groups can also be divided into halves according to the principle of purines on one side, pyrimidines on the other (and for a completely heterogeneous group there are two natural ways to do this), but neither for halves of each group, nor even when trying to combine these groups, the purine-pyrimidine division leads either to the Shcherbak–Makukov numbers or to the equilibrium of the weights in the halves.

Table 6. The weight of the side chains of the codons in which one pair of identical bases is a pyrimidine and the other is a different base

TTC **F** 91	TTA **L** 57	TTG **L** 57	CCT **P** 41	CCA **P** 41	CCG **P** 41
CTT **L** 57	ATT **I** 57	GTT **V** 43	TCC **S** 31	ACC **T** 45	GCC **A** 51
TCT **S** 31	TAT **Y** 107	TGT **C** 47	CTC **L** 57	CAC **H** 81	CGC **R** 100
∑ = 999					

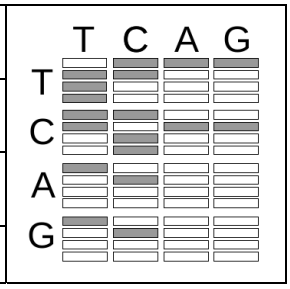
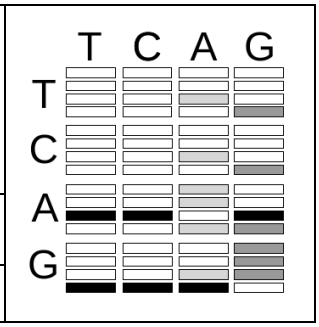


Table 7. The weight of the side chains of the codons in which one pair of identical bases is a purine and the other is a different base. Different filling colors on the genetic code diagram correspond to different subgroups of codons in the table separated by a double line

AAT			GGT **G** 1	GGC **G** 1	GGA **G** 1
N 58	AAC **N** 58	AAG **K** 72	TGG **W** 130	CGG **R** 100	AGG **R** 100
TAA **stop** 0	CAA **Q** 72	GAA **E** 73			
ATA I 57	ACA T 45	AGA R 100	GTG V 43	GCC A 15	GAG E 73
∑ = 999					



Shcherbak and Makukov showed that there is a variant of splitting completely homogeneous and completely heterogeneous groups in half, when some SM-numbers are obtained for the sums, but this is a special choice among many otherwise equal variants, not based on any *a priori* clear rule (like purines on one side, pyrimidines on the other), so this result may be considered as a result of fitting. We do not present it here.

Signature 3.3. Triple Symmetry of Codons with Two Identical Bases Which Are Purines

A group of codons in which a pair of identical bases are purines (with weight 999, see Signature 3.2. and Table 7) can be divided into three equal-size subgroups according to the following principle: the first group contains two bases A in a row, the second group contains two bases G in a row, the third group contains either two bases A separated by another base or two bases G separated by another base. These three groups are separated by double lines in Table 7. It turns out that the sum of the masses of the side chains of each subgroup is equal to the SM-number 333. In other words, there is a triple equilibrium of the subgroups, and the equilibrium number is an SM-number.

Signature 3.4. Codons with Two Identical Bases and a Single Purine or a Single Pyrimidine

If, in a group of codons with two identical bases and a different third base (as in Signature 3.2.), we consider separately codons with single purine and single pyrimidine, it turns out that the corresponding sums of the side chain weights are SM-numbers separately: for single purine 888 (see Table 8); for single pyrimidine $1 \times 999 + 111$ (see Table 9). The independent signature here

is that the weight of one of the subgroups is an SM-number, since the fact that the second subgroup is also an SM-number follows from the fact that the sum of the weights of both subgroups is an SM-number (according to Signature 3.2.).

Table 8. Two identical bases and a different third single purine

TTC **F** 91	CCT **P** 41	AAT **N** 58	AAC **N** 58	GGT **G** 1	GGC **G** 1
CTT **L** 57	TCC **S** 31	TAA **stop** 0	CAA **Q** 72	TGG **W** 130	CGG **R** 100
TCT **S** 31	CTC **L** 57	ATA **L** 57	ACA **T** 45	GTG **V** 43	GCG **A** 15
∑ = 888					

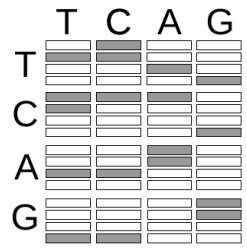
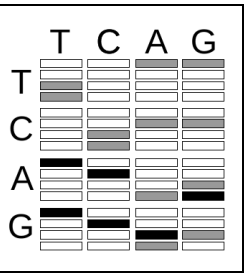


Table 9 (for single pyrimidines) is constructed according to the following simple symmetry principle. The first row lists all possible codons when the single pyrimidine is in the last position and the first two bases are the same. The next two rows are obtained by cyclic clockwise rearrangement of the bases in the first row. This exhausts all possible codons with two identical bases and a single pyrimidine. It turns out that in this case the sum of the masses of the second row of the table is the SM-number 333, and this is a new independent SM-signature. Since the total weight for the whole of Table 9 is the SM-number 999 × 1 + 111, the sum of the first and last rows also turns out to be an SM-number (it is 777), which is no longer an independent signature.

Table 9. Two identical bases and a different third single pyrimidine. The group of codons with the sum of weights 333 is highlighted in black on the genetic code diagram

TTA **L** 57	TTG **L** 57	CCA **P** 41	CCG **P** 41	AAG **K** 72	GGA **G** 1	∑ = 269
ATT **I** 57	GTT **V** 43	ACC **T** 45	GCC **A** 15	GAA **E** 73	AGG **R** 100	∑ = 333
TAT **Y** 107	TGT **C** 47	CAC **H** 81	CGC **R** 100	AGA **R** 100	GAG **E** 73	∑ = 508
∑ = 999 × 1 + 111 = 333 + 777						



Signature 3.5. Total Weights of the Codons with the First Pyrimidine or the First Purine

Calculating the sum of the total weights of the coding products corresponding to the codons with the first pyrimidine, without taking into account the de-

gree of degeneracy in the columns and assuming that the weight of the stop signal is equal to 74, gives the SM-number 1 × 999 + 777 (see Table 10).

A similar calculation of the sum of the total weights of the coding products corresponding to codons with the first purine gives 1517 (see Table 11). The number 1517 is a multiple of 37, which is nontrivial, although 1517 is not an SM-number: 1517 = 1 × 999 + 518.

These two signatures are missing in the article (Shcherbak and Makukov 2013) where no signature was found for the coding products with the first purine. For the coding products with the first pyrimidine, the sum of the side chain weights is found to be 814 (divided by 37), and the sum of the masses of the constant parts is calculated, which is again 814, that is weight equilibrium is observed. This last calculation, however, assumes that the total weight of the **stop** signal has weight zero, which, as we have explained, is not entirely meaningful if at the same time we want to consider the weight of the side chain of the **stop** signal to be zero. In other words, the Shcherbak–Makukov signature does not fit well into the accepted logic.

Table 10. Total weight of the coding products corresponding to codons with the first pyrimidine

TTY **F** 165	TTR **L** 131	TCN **S** 105	TAY **Y** 181	TAR **stop** 74	TGY **C** 121	→ →
TGA **stop** 74	TGG **W** 204	CTN **L** 131	CCN **P** 115	CAY **H** 155	CAR **Q** 146	CGN **R** 174
$\sum = 1 \times 999 + 777$						

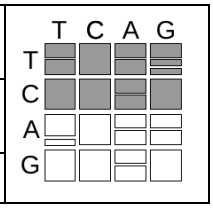
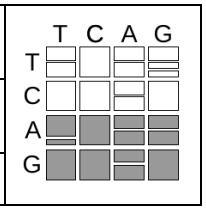


Table 11. Total weight of the coding products corresponding to the codons with the first purine

ATH **I** 131	ATG **M** 149	ACN **T** 119	AAY **N** 132	AAR **K** 146	AGY **S** 105	→ →
AGR **R** 174	GTN **V** 117	GCN **A** 89	GAY **D** 133	GAR **E** 147	GGN **G** 75	
$\sum = 1517 = 41 \times 37$						



Signature 3.6. Decomposition of Codons

Shcherbak and Makukov understand codon decomposition as the following procedure (Shcherbak and Makukov 2013). Each codon is mentally decomposed into its constituent bases, after which each base is assigned the weight of the product encoded by the original codon, 'decomposition' of which resulted

in the given single base. The weight of the product can be the weight of the side chain, the total weight, or the weight of the constant part of the amino acid in different cases. Using such a decomposition of triplets, Shcherbak and Makukov found the following signatures.

The sum of the weights of the side chains on all single bases of the entire table of the genetic code, calculated this time taking into account the degree of degeneracy, turns out to be the SM-number $10 \times 999 + 222$. It should be noted that this signature is not independent of the signatures found above. Indeed, we already have the full masses of the octets assuming the weight of the **stop** signal is equal to 74: 3700 and 4440 for the first and second octets. Thus, for the total weight of the entire table, we obtain $3700 + 4440 = 8140$. Since the weight of each octet is divided by 74, the sum of the masses is again divided by 74. From here it is easy to find the sum of the masses of the side chains. To do this, subtract weight of the constant part from 8140 exactly as many times as there are cells in the table: $8140 - 74 \times 64 = 3404$. The resulting number is again divided by 74, therefore it is also divided by 37. To get the weight of side chains on all bases, it is necessary to multiply the weight of all side chains on codons by three, because each codon is counted exactly three times when counting on separate bases. Therefore, the result will necessarily be divisible by $37 \times 3 = 111$, which means that it will necessarily be an SM-number. Indeed, it is easy to check: $3404 \times 3 = 10 \times 999 + 222$. That is, although Shcherbak and Makukov found this signature to be independent, it is actually derived from the signatures we obtained earlier (beginning of Section 5), which Shcherbak and Makukov did not know.

However, another signature for codon decomposition found by Shcherbak and Makukov turns out to be independent and non-trivial. It turns out that the sum of the side chain masses only for base T separately is the SM-number $2 \times 999 + 666$. It follows that the sum of the side chain masses calculated for the bases C, G, A is also an SM-number, since the difference of any two SM-numbers is again an SM-number (in this case this number is $7 \times 999 + 555$). Thus, the method of codon decomposition actually leads to one independent SM-signature.

Signature 3.7. Rumer Pairs: M = (T, G) and K = (C, A)

The Rumer transformation $R = (T \leftrightarrow G, C \leftrightarrow A)$ distinguishes two base pairs M = (T, G) and K = (C, A) in which the bases swap places. Shcherbak and Makukov found signatures associated with these two particular base pairs.

If all columns of the genetic code table are divided into two halves according to whether the first base belongs to pair M or pair K, and if we calculate for each half the sum of the side chain weights, counting the coding products without taking into account the degeneracy in the columns (Shcherbak–Makukov rule), the weight of half M is 654 (see Table 12) and the weight of half K is 789 (see Table 13).

These numbers are not interesting in themselves, they do not even divide by 37. The sum 789 + 654 = 1 × 999 + 444 is, of course, an SM-number, but it is not an independent signature, as it follows from Signature 3.1. But if the code table is divided into two halves in a similar way, but now using the middle codons of the columns, the weight for group M turns out to be 789 (see Table 14), and for group K – 654 (see Table 15). The weights of the groups turn out to be the same with permutation accuracy, and this fact is algebraically independent of other known signatures.

This is not all. If we assign to one group the columns in which both first codons belong to M and to another – the columns in which both first codons belong to K, the weights of the groups turn out to be the same, both equal to 369, and this fact is also algebraically independent (see Table 16). There is a balance 369 = 369.

Table 12. Sum of the side chain masses of codons with the first base belonging to the pair M = (T, G)

TTY **F** 91	TTR **L** 57	TCN **S** 31	TAY **Y** 107	TAR **stop** 0	TGY **C** 47	→ →
TGA **stop** 0	TGG **W** 130	GTN **V** 43	GCV **A** 15	GAY **D** 59	GAR **E** 73	GGN **G** 1
∑ = 654						

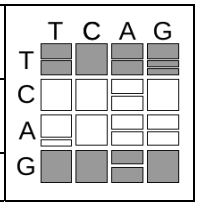


Table 13. Sum of the side chain masses of codons with the first base belonging to the pair K = (C, A)

CTN **L** 57	CCN **P** 41	CAY **H** 81	CAR **Q** 72	CGN **R** 100	ATH **I** 57	→ →
ATG **M** 75	ACN **T** 45	AAY **N** 58	AAR **K** 72	AGY **S** 31	AGR **R** 100	
∑ = 789						

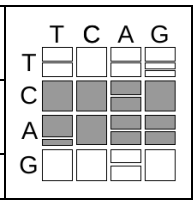


Table 14. Sum of side chain masses of codons with the second base belonging to the pair M

TTY **F** 91	TTR **L** 57	CTN **L** 57	ATH **I** 57	ATG **M** 75	CTN **V** 43	→ →
TGY **C** 47	TGA **stop** 0	TGG **W** 130	CGN **R** 100	AGY **S** 31	AGR **R** 100	GGN **G** 1
∑ = 789						

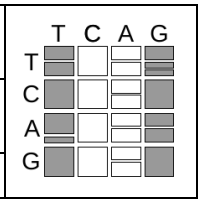


Table 15. Sum of side chain masses of codons with the second base belonging to the pair K

TCN S 31	CCN P 41	ACN T 45	GCN A 15	TAY Y 107	TAR stop 0	→ →
CAY H 81	CAR Q 72	AAY N 58	AAR K 72	GAY D 59	GAR E 73	
∑ = 654						

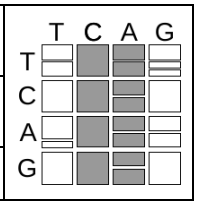
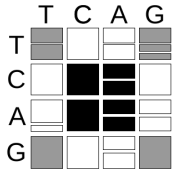


Table 16. Weights of groups of coding products with both the first codons belonging to the M or K groups of bases

Both first bases in M			
TTY **F** 91	TTR **L** 57	TGY **C** 47	TTA **stop** 0
TGG **W** 130	GTN **V** 43	GGN **G** 1	
∑ = 369			

Both first bases in K		
CCN **P** 41	CAY **H** 81	CAR **Q** 72
ACN **T** 45	AAY **N** 58	AAR **K** 72
∑ = 369		



6. The Third Information Level + TGA-switch

A ‘symmetrized’ genetic code is considered in the paper by Shcherbak and Makukov (2013) together with the universal genetic code. The symmetrization operation consists in the redefinition of the function of the single TGA codon from signal **stop** coding to cysteine (**C**) coding:

$$\text{TGA} : \textbf{stop} \rightarrow \text{TGA} : \textbf{C}. \qquad (6)$$

As a result of this symmetrization, the AT and TG columns acquire the same structure (see Fig. 1), and a number of new strong information signatures are detected in the modified genetic code (see below). It turns out that such a symmetrized genetic code exists in nature: it is one of the alternative genetic codes. It is possessed by the single-celled creature *euplotidium* from a genus of ciliates. The alternative genetic codes in the context of the issues of this article are discussed in Section 9.3.

In fact, the logic of the symmetrization operation is close to the logic of the ‘activation key’ in redefining the proline structure. However, the similarity of the logical structure of both operations is not noticed in the article by Shcherbak and Makukov (2013), and the term ‘activation key’ is not used in connection with the symmetrization of the universal code. The resulting code structure has the same virtual character as the weight 74 of the proline side chain after redefinition of the proline structure, despite the fact that the resulting code is

found in nature as an alternative genetic code. Alternative genetic codes are most likely produced by later mutations of the basic universal genetic code, so if we are discussing the probable artificial nature of information signatures of the genetic code, we should look for them in the original version of the code, where they could have been introduced. The coincidence of the symmetrized version of the code with the version of the *euplotidium* code is almost likely a mere coincidence.[5]

At the first sight, the activation key in the form of symmetrization of the code (6) is absolutely inadmissible in the logic of the sequence of information levels adopted in present paper. In fact, the transformation (6) does not affect the first octet of Rumer and therefore does not change its total weight 3700, but the weight of the second octet becomes either 4265, if we consider the total weight of the product **stop** to be zero, or 5449, if we consider the total weight of the product **stop** to be 74, as was adopted at the beginning of the discussion of the third information level. Neither of the numbers 4265 and 5449 is divisible by 74, much less by 740, so the transformation (6) completely destroys the second information level (a very simple level of 'attracting attention') and all subsequent evidence for the special role of the pair (10, 37) and the number 111. It seems that this completely destroys the logic leading from the first information level with Rumer's symmetries to the third, more complex level through the second information level.

However, another way of looking at the transformation (6) is possible. This transformation can be seen not as an activation key that leads to new information signatures by destroying the second information layer, but as a two-position switch:

$$\text{TGA} : \textbf{stop} \leftrightarrow \text{TGA} : \textbf{C}. \tag{7}$$

In its initial position TGA : **stop**, the TGA-switch turns on the second information level but turns off all signatures related to the symmetrization of the code, and in position TGA : **C** it turns off the second information level but turns on the signatures related to the symmetrization. Thus, one and the same text of the genetic code allows, due to the TGA-switch, two variants of reading, which coexist in the genetic code despite the fact that they are mutually exclusive.

Note that the TGA-switch in the position TGA : **C** does not violate the above signatures of the third information level due to the special Shcherbak–Makukov summation rule used in the construction of third-level signatures. The only exception is the Signature 3.6. (Decomposition of codons). This interest-

[5] The possibility that both variants of the genetic code – standard and *euplotidium* – were present from the very beginning of life on Earth on an equal footing was mentioned in a small section 'Two versions of the code' in the paper by Shcherbak and Makukov (2013). In this case, however, one would expect that a comparable number of species would have used both types of genetic code now. This is not the case, so this hypothesis seems very unlikely.

ing fact allowed Shcherbak and Makukov to explain the possibility of using together with the basic version of the universal genetic code also an alternative symmetrized version of the code in the following way (Makukov and Shcherbak 2018). Since the symmetrization does not destroy the basic numerical signatures of the *universal code* (except for one), but introduces new very strong symmetries and new numerical signatures, the symmetrization procedure effectively increases the information capacity of the universal genetic code and, in fact, can be considered itself as a part of the information record embedded in *the universal genetic code*. This treatment looks very reasonable, and our proposed treatment of the symmetrization operation as a switch extends the Shcherbak–Makukov treatment.

Now, according to Shcherbak and Makukov (2013), we will present the information signatures of the genetic code resulting from the symmetrization in the switch position TGA : **C**. Until the end of the current section, we consider only the symmetrized table of the genetic code.

If we examine the columns of the genetic code table, we can find four degrees of degeneracy of the coding product within the columns. Each column of the first octet corresponds to degeneracy four, which we will denote by the Roman numeral **IV**. In the columns of the second octet we find the coding degeneracy from three to one. The degeneracy one occurs in the singletons ATG : **M** and TGG : **W**, degeneracy three occurs in the triplets TGH : **C** and ATH : **I**, and there are many doublets with coding degeneracy two: TAR : **stop** and others. Thus, all the second octet coding products can be divided into three degeneracy groups: **I**, **II**, **III**.

Table 17 represents the structure called an *ideogram* in the paper by Shcherbak and Makukov (2013). In the table, the degeneracy groups from **I** to **III** are arranged in descending order, and within each group the coding products are arranged in ascending order of side chain weights. For each product, the side chain weight and the codons encoding that product are given. It turns out that the ideogram has a number of remarkable symmetries and exhibits several numerical signatures leading to SM-numbers. Below, all these symmetries and signatures are considered in turn. The signatures of the symmetrized code belong to the third information level, but their numbers will be marked with the suffix B, unlike the third level signatures of the original genetic code.

Table 17. The ideogram of the second octet of the symmetrized genetic code (see text for explanation)

group	**III**		**II**												**I**	
product	**C**	**I**	**stop**	**S**	**L**	**N**	**D**	**Q**	**K**	**E**	**H**	**F**	**R**	**Y**	**M**	**W**
M_{side}	47	57	0	31	57	58	59	72	72	73	81	91	100	107	74	130
base 1	**T**	**A**	**T**	**A**	**T**	**A**	**G**	**C**	**A**	**G**	**C**	**T**	**A**	**T**	**A**	**T**
base 2	G	T	A	G	T	A	A	A	A	A	A	T	G	A	T	G
base 3	H	H	R	Y	R	Y	Y	R	R	R	Y	Y	Y	Y	G	G

Table 18. Symmetry of the second bases of the first octet and the central part of the second octet

Octet **I**, group **IV**	product	G	A	S	P	V	T	L	R
	M_{side}	1	15	31	41	43	45	57	100
	base 2	R	Y	Y	Y	Y	Y	Y	R
Octet **II**, group **II**	base 2	Y	R	R	R	R	R	R	Y
	M_{side}	57	58	59	72	72	73	81	91
	product	L	N	D	Q	K	E	H	F

Signature 3.1B. Symmetries of the Base 1 Line of the Ideogram

The first and last five bases in the base 1 string are symmetric about the center of the string. These bases are marked in light gray and form the string TATAT which itself is mirror symmetric. The central AGC|AGC group (marked in dark gray) consists of two identical AGC subgroups (or has exact shift symmetry with respect to the center).

Signature 3.2B. Symmetry of the Base 2 Line of the Ideogram

Base 2 line of the ideogram is exactly mirror symmetric.

Signature 3.3B. The Symmetry between the Degeneracy Groups of the First and Second Octets

Let the octet **I**, written in ascending order of amino acid weight, is centered with the center of the ideogram Table 17 (octet **II**). In each case, we take the rows of the middle bases (base 2) and write their types instead of the bases themselves – purine (R) or pyrimidine (Y). This construction is shown in Table 18. Then we find that the resulting strings of bases of the octets **I** and **II** are, first, identical up to the inversion, and, second, each of the strings is mirror-symmetric in itself (see the two middle strings of Table 18).

Signature 3.4B. Symmetries of the Virtual Codons

The exactly mirror symmetric string base 2 (see Table 17) has a number of remarkable properties besides this symmetry itself.

First, let us note that the coding product methionine (**M**) plays another role in the genetic code besides being an amino acid. Methionine is the starting point of practically every protein that is synthesized. Therefore, the ATG codon

means not only methionine, but also the **start** command, which is the opposite of the **stop** command.

Table 19. The ideogram of the second octet of the symmetrized genetic code and the virtual codons (see text for explanation)

group	**III**		**II**												**I**		∑
product	**C**	**I**	**stop**	**S**	**L**	**N**	**D**	**Q**	**K**	**E**	**H**	**F**	**R**	**Y**	**M**	**W**	
M_{side}	47	57	0	31	57	58	59	72	72	73	81	91	100	107	75	130	
base 2	G	T	A	G	T	A	A	A	A	A	A	T	G	A	T	G	
v-product	–	**stop**		**stop**		**K**		**start/M**			**start/M**						
M_{side}	–	0		0		72		75			75						222
base 2	G	T	A	G	T	A	A	A	A	A	A	T	G	A	T	G	
v-product	–	S		K		K		C			–						
M_{side}	–	31		72		72		47			–						222

Starting from the first occurrence of the TAG base sequence in the base 2 string (marked in dark gray in Table 19), we divide the string into triplets until the end of the string, as shown in Table 19. Now let us see what coding products these triplets would correspond to if they were real codons. It turns out that the sequence of these 'virtual codons' corresponds to a symmetric string of 'virtual coding products' (v-products):

stop → stop → K → start → start.

Furthermore, the sum of the side chain weights of these virtual products turns out to be SM-number 222 (see Table 19).

If we move the starting position of the reading frame one step to the right, we will get the sequence of 'virtual codon products' **S → K → K → C**, where the sum of the side-chain weights is again 222 (see the last three lines of Table 19).

So not only two new independent SM-numbers appear here, but there is also an equilibrium between them.

7. The Fourth Information Level. Masses of the Chain Links of Amino Acid Molecules

The fourth information level of the genetic code signatures is related to the use of a new resource. If we look at the structure of the side chains of amino acids (Shcherbak and Makukov 2013; McMurry 2023), we will see that all of them are more or less long chains whose links are based on heavy atoms, which

in different cases can be carbon, oxygen, nitrogen or sulfur, linked to a certain number of hydrogen atoms (from zero to three). The only exception is glycine, whose entire side chain consists of a single hydrogen atom and contains no heavy atoms. The successive links of these chains can be uniquely renumbered, and Shcherbak and Makukov use the concept of a level[6] for this purpose, and the levels are numbered with Greek letters starting with beta: *β, γ, δ, ε, ζ, …* The *α* value is assigned to the constant part of the amino acid, the *β* level is the closest to the constant part, and so on. Glycine has a single *β* level consisting of a hydrogen atom. Some complication to this picture is that the chains are sometimes split into two parallel subchains, but this does not prevent the 'parallel' links of the split part from uniquely assembling into a single level. In the article by Shcherbak and Makukov (2013), the examples of the splitting of amino acid side chains into levels are shown in Figs 7b and 8a. The new resource allows the identification of several new signatures. All signatures of the fourth information level below were found by Shcherbak and Makukov. For the **stop** signal, it is natural to consider the masses of all levels to be zero, consistent with the fact that the weight of the entire **stop** side chain is assumed to be zero.

Signature 4.1. Level β of the First Octet

Shcherbak and Makukov found that the sum of the masses of the links of level *β* for the whole first octet, calculated without taking degeneracy into account, is the SM-number 111 (see Table 20). Taking degeneracy into account, the result would be 444. There is no *a priori* reason to expect that the sum of the masses calculated by any method for any set of amino acids will turn out to be an SM-number or at least a multiple of 37. The signature found, like all other signatures of the fourth information level, is independent.

Table 20. Masses of the links of level β for the octet **I**

amino acid	**G**	**A**	**S**	**P**	**V**	**T**	**L**	**R**	T C A G T C A G
level *β* weight	H 1	CH_3 15	CH_2 14	C_2H_3 27	CH 13	CH 13	CH_2 14	CH_2 14	
∑	111								

Signature 4.2. Levels *β* and *γ* of the Second Octet

The sum of the masses of the levels *β* and *γ* of the coding products of the entire second octet, without considering the degeneracy in the columns, is the SM-number 444 (see Table 21).

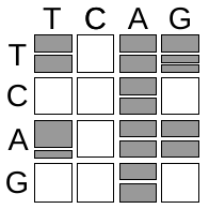
[6] Same as the concept of a 'locant' in the nomenclature of organic chemistry.

Table 21. Sum of the masses of levels β and γ of the whole octet **II**

amino acid	**F**	**L**	**Y**	**stop**	**C**	**stop**	**W**	**H**	→
level β weight	CH_2 14	CH_2 14	CH_2 14	– 0	CH_2 14	– 0	CH_2 14	CH_2 14	→
level γ weight	C 12	CH 13	C 12	– 0	SH 33	– 0	CH 13	C 12	→

amino acid	**Q**	**I**	**M**	**N**	**K**	**S**	**R**	**D**	**E**
level β weight	CH_2 14	CH 13	CH_2 14	CH_2 14	CH_2 14	CH_2 14	CH_2 14	CH_2 14	CH_2 14
level γ weight	CH_2 14	C_2H_5 29	CH_2 14	C 12	C 12	OH 17	CH_2 14	C 12	CH_2 14

∑	444



Signature 4.3. Codons with the First Pyrimidine or the First Purine

The codon count is exactly the same as in Signature 3.5. of the third information level here. The coding products of codons with the first pyrimidine or the first purine are counted separately, without taking into account the degree of degeneration in the columns. Then, for codons with the first pyrimidine the sum of the weights of the β and γ levels gives the SM-number 333 (see Table 22), and for codons with the first purine the sum of the weights of the β, δ, ζ levels again gives the SM-number 333 (see Table 23).

Table 22. Codons with the first pyrimidine, sum of the weights for β and γ levels. Level weights are shown in parentheses

amino acid	TTY **F**	TTR **L**	TCN **S**	TAY **Y**	TAR **stop**	TGY C	→
level β	CH_2 (14)	CH_2 (14)	CH_2 (14)	CH_2 (14)	– (0)	CH_2 (14)	→
level γ	C (12)	CH (13)	OH (17)	C (12)	– (0)	SH_2 (33)	→
amino acid	TGA **stop**	TGG **W**	CTN **L**	CCN **P**	CAY **H**	CAR Q	CGR R
level β	– (0)	CH_2 (14)	CH_2 (14)	C_2H_3 (27)	CH_2 (14)	CH_2 (14)	CH_2 (14)
level γ	– (0)	C (12)	CH (13)	CH_2 (14)	C (12)	C (12)	CH_2 (14)
∑	333						

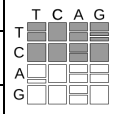


Table 23. Codons with the first purine, sum of weights for β, δ and ζ levels. Level weights are shown in parentheses

amino acid	ATH **I**	ATG **M**	ACN **T**	AAY **N**	AAR **K**	AGY **S**	→	
level β	CH (13)	CH_2 (14)	CH (13)	CH_2 (14)	CH_2 (14)	CH_2 (14)	→	
level δ	CH_3 (15)	S (32)	− (0)	ONH_2 (32)	CH_2 (14)	− (0)	→	
level ζ	− (0)	− (0)	− (0)	− (0)	NH_2 (16)	− (0)	→	T C A G T C A G
amino acid	AGR **R**	GTN **V**	GCN **V**	GAY **D**	GAR **E**	GGN G		
level β	CH_2 (14)	CH (13)	CH_3 (15)	CH_2 (14)	CH_2 (14)	H (1)		
level δ	CH_2 (14)	− (0)	− (0)	O_2H (33)	C (12)	− (0)		
level ζ	C (12)	− (0)	− (0)	− (0)	− (0)	− (0)		
∑	333							

Here we can note some elements of adjustment in deriving these two sums in the sense that the choice of levels to count is rather *ad hoc*. Other combinations do not lead to interesting results. However, this choice does have some special symmetry: for pyrimidines the levels β and γ strictly go in a row, and for purines the levels β, δ, ζ go strictly through one, not to mention that the sums in both cases turn out to be really quite unexpectedly the same.

8. The Fifth Information Level. ARS Symmetries

Genetic information is passed from generation to generation by being written on the DNA substrate. But the way DNA information is decoded – the genetic code – is also passed from generation to generation. The genetic code is a complex information structure, and in order to be transmitted from generation to generation, it must also be written on something. Information cannot be transmitted without being recorded on some medium. What is the medium on which the genetic code is written? Obviously, this medium itself is an integral part of the realization of the genetic code in nature, just as genetic information is inconceivable without its medium – DNA molecules.

The genetic code is realized with the help of a specific set of tRNA molecules (transport RNAs) and ARS molecules (aminoacyl-tRNA synthetase), which recode the information of the DNA codons into a specific nomenclature of amino acids. This specific set of tRNA and ARS molecules is passed from generation to generation during cell division, and it is the medium – the carrier of information – on which the genetic code is written. Consequently, if we are

interested in information structures related to the genetic code, the medium of the genetic code record should not be left aside. The transition to the analysis of possible information signatures associated with the substrate of the genetic code record clearly marks the transition to a new, now fifth, information level of genetic code signatures (this analysis is completely new; it is absent in the articles of Shcherbak and Makukov).

The ARS molecule binds to multiple sites on tRNA and amino acid molecules, which is called superspecificity of mutual recognition. All ARS are divided into two classes, based on structural similarity and the way they aminoacylate tRNA. Class 1 ARS (ARS-1) are mostly monomers. ARS-2 are mostly dimers. In addition, ARS-1 recognize 'their' tRNA from the so-called 'small groove' side of the acceptor mini-helix of ribosomal RNA, while ARS-2 recognize 'their' tRNA from the 'large groove' side.

Ten amino acids correspond to a single ARS-1 for each amino acid, another nine amino acids correspond to a single ARS-2 for each amino acid, and a single amino acid, lysine, corresponds to one ARS-1 and one ARS-2. Lysine is a strange exception among all the other amino acids, because lysine does not correspond to one ARS molecule of a certain type, but to two at once, and of different types. It would seem natural to divide all the amino acids into two groups according to the classes of ARS to which they belong, but this turns out to be impossible because of the ambiguity associated with lysine. A possible solution to this problem is similar to the use of a proline activation key. We artificially assign ARS-2 to lysine, then all amino acids are divided into two equal groups of ten amino acids each, belonging to either ARS-1 or ARS-2.

It should be noted that there are probably hints of a special role of lysine in the information signatures of the genetic code. In addition to the special relation to ARS classes, we can note the double central position of lysine (more precisely, the AAA codon) in the fully symmetric ideogram line (see Table 19), and its central position in both lines of the virtual products (**stop → stop → K → start → start**) and (**S → K → K → C**). There are some other signs of a special role of lysine, but we do not consider them here. What that special role might be is not yet clear.

The division of ARS into two equal classes is already an indication of the existence of symmetry associated with ARS molecules, just as two octets of codons are associated with Rumer symmetry. This leads us to take a closer look at ARS.

The presence of unusual formal features of ARS is revealed when the object of analysis is the nucleotides and amino acids that characterize these intermediates. In Table 24, the amino acids are divided into two groups according to their associated ARS, and within each group the amino acids are ordered by increasing weight (side chain weight or total weight, irrelevant here). The first

bases of the corresponding codons are shown together with the amino acids (in bold). The bases corresponding to the homogeneous columns of the genetic code table (degeneracy level in columns **IV**) are given for the 'super-degenerated' amino acids **S**, **L**, and **R** (full degeneracy level **VI**).

Table 24. Two classes of ARS represented by their specific amino acids with the first letters of the codons (see text). Arrows indicate increasing masses of amino acids

→					
ARS-1	**V** G	**C** T	**L** C	**I** A	**Q** C
	E G	**M** A	**R** C	**Y** T	**W** T
G G	**A** G	**S** T	**P** C	**T** A	ARS-2
N A	**D** G	**K** A	**H** C	**F** T	
→					

When the positions of the rows of both classes are shifted by one column relative to each other, as shown in Table 24, the symmetry of the first bases of the codons associated with ARS is revealed. The G and C nucleotide positions of the central four columns of the table appear to be assembled in separate columns (black background), while the other two columns, with bases A and T, appear to be symmetric with respect to simultaneous vertical inversion and (separately) with respect to cyclic shift (dark gray background). The first four nucleotides of the two outer half-columns (light gray background) together form the complete chain CTGA. The common colors light gray and white in rows indicate the complete chains of the first four bases GTCA and GACT, respectively, which appear horizontally in each class.

Note that amino acids **L** and **I** have the same total weight 131 and both belong to class ARS-1, so their order in Table 24 is, strictly speaking, undefined. However, it is precisely the order **L**, **I** that leads to the symmetry of Table 24, so we can say that the arbitrariness in the choice of the order is removed *ex post facto*. Note that the ordering of **L** and **I** only affects the symmetry of Table 24, but not other symmetries and information signatures related to ARS, which will be discussed below.

The symmetry of Table 24 looks quite strange. As we can see, the idea to go deeper in the search for information signatures associated with ARS, which initially arose simply from the presence of two equal groups of ARS of the first and second type, is reinforced by the new symmetries of Table 24. The picture

begins to look more and more like the Rumer symmetry situation. Is it possible to go even deeper in the analysis of the ARS structure, similar to what was done on the second information level of the genetic code using amino acid weights?

Table 25. Amino acid numbering corresponding to ARS-1 and ARS-2 (see text). The weight accretion of amino acids of each ARS class is indicated by the intensity of the gray color of the corresponding cell

V	**C**	**L**	**I**	**Q**	**E**	**M**	**R**	**Y**	**W**	ARS-1
1	2	3	4	5	6	7	8	9	10	+
G	**A**	**S**	**P**	**T**	**N**	**D**	**K**	**H**	**F**	ARS-2
−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	−

The complexity and irregularity of ARS molecules do not allow us to use their nucleon weight as a basis for numerical analysis. Instead, we take a different approach. We label the amino acids specific to each ARS with ordinal numbers according to the increasing weight of the amino acid. Since the ARS classes are equal in this respect and none has an advantage over the other, the numbering of the amino acids of each class should emphasize this and at the same time distinguish the classes. A natural solution to this problem, based on the symmetry of the ARS classes, is to label the amino acids corresponding to each class with positive and negative numbers (which have no independent meaning) according to the increasing values of their masses, *i.e.* the ARS-2 class is from −10 to −1, and ARS-1 from +1 to +10. This approach is shown in Table 25. The order of the amino acids **L**, **I** in the Table 25 is the same as in the Table 24.

Let us do the following construction. For each of the bases C, T, A, G (the bases are written in order of increasing molecular weight), we will write horizontally, in order of increasing weight, the amino acids encoded by triplets in which the corresponding base is the first, using the same rule as in the construction of Table 24: the 'super-degenerated' amino acids **S**, **L**, and **R** correspond to the bases from the homogeneous columns of the genetic code table (column degeneration level **IV**). Let us also write in the corresponding cells of the table the indices of the amino acids from Table 25. We obtain Table 26, which we will call Calligram-B.

Table 26. Calligram-B. For each base C, T, A, G (following in increasing order of weight represented by gray level), the amino acids corresponding to the triplets with this first base are written out in increasing order of weight and numbered according to their position in the sequence corresponding to the ARS-1 (+, black text) and ARS-2 (—, gray text) classes (see Table 25)

C	–7, **P**	+3, **L**	+5, **Q**	–2, **H**	+8, R	(3, 2)
T	–8, **S**	+2, **C**	–1, **F**	+9, **Y**	+10, W	(3, 2)
A	–6, **T**	+4, **I**	–5, **N**	—3, **K**	+7, M	(2, 3)
G	–10, **G**	–9, **A**	+1, **V**	—4, **D**	+6, E	(2, 3)
∑	–31	0	0	0	+31	
	(0, 4)	(3, 1)	(2, 2)	(1, 3)	(4, 0)	

The numbering of the amino acids according to their weight in the ARS classes leads to a very peculiar position of them in the Calligram-B (see Table 26). The sums of the numbers in the columns of this calligram (the row 'Σ' of the table) are antisymmetric with respect to the central column. The sums of the numbers in columns 2, 3, 4 are exactly equal to zero, and the sums of the edge columns are equal in modulo to 31 but opposite in sign. Note that this signature does not depend on the choice of the order of **L**, **I** in Table 25.

There are two more symmetries in the Calligram-B. If we look at the number of occurrences of ARS-1 and ARS-2 in each column of Calligram-B, we will get the fully symmetric sequence (0,4), (3,1), (2,2), (1,3), (4,0) (see Table 26). For rows, the fully symmetric sequence (3,2), (3,2), (2,3), (2,3) is obtained again in a similar way. There is no reason to expect such symmetry in general case.

We can estimate the scale of probability of realization of symmetries that take place in Calligram-B. Let us estimate the probability of the following features: 1) antisymmetry in the distribution of sums of numbers by columns, P_{as} (similar to row −31, 0, 0, 0, +31); 2) symmetry in the distribution of ARS types by columns, $P_{col.sym}$ (similar to row (0,4), (3,1), (2,2), (1,3), (4,0)); 3) symmetry in the distribution of ARS types by rows, $P_{row.sym}$ (similar to row (3,2), (3,2), (2,3), (2,3)). Monte Carlo calculations provide the probabilities separately: P_{as} = 1.11 %, $P_{col.sym}$ = 18.8 %, $P_{row.sym}$ = 34.4 %. All three symmetries occur simultaneously, as in the real Calligram-B, with probability only P_{joint} = 0.378 %. In other words, the observed Calligram-B symmetry is rather unlikely.

The surprising number of symmetries of Calligram-B is very reminiscent of the 'first level of attention' associated with Rumer symmetries, and looks like an invitation to further explore ARS structures. This problem has not yet been solved.

9. Discussion

Thus, in the genetic code and in the molecular structures directly related to it, a considerable number of information structures are found that have a rather special appearance and do not look 'random' in a certain sense. This, of course, may suggest their artificial origin.

However, is the task of triggering the evolution of life with artificial information in the genetic code, even on the scale of an entire planetary system, so fantastic and unfeasible that it is a strong argument against the artificial nature of the information signatures of the genetic code?

Note here that the introduction of information patterns into the genetic code does not require the creation of life 'from scratch', but can be accomplished by modifying the already existing genetic code (*e.g.,* the genetic code of the 'creator' itself). Such an operation can be accomplished using already existing terrestrial technologies. For example, Hohsaka *et al.* (2001) were able to transform some codons from three-nucleotide to four- and even five-nucleotide codons. Hoesl *et al.* (2015) were able to change the value of TGG codons in the bacterium Escherichia coli from tryptophan to thienopyrrole-alanine, which is not found in nature. There are many other examples of successful artificial intervention in the genetic code, suggesting that it can be artificially modified without destroying other mechanisms of cell operation (some details can be found in Shcherbak and Makukov 2013; Makukov and Shcherbak 2018). Bacteria with a modified genetic code containing artificial information patterns can be placed in a frozen state in containers sent to the formation region of a young planetary system. Bacteria can survive in a frozen state for hundreds of thousands or even millions of years, so sublight speed of flight would not be required to deliver such containers to other stars. The containers of frozen bacteria can be programmed to continuously contaminate the environment of the protoplanetary cloud. Once the protoplanetary cloud contains the first bacterial habitat, the bacteria will immediately begin to reproduce, and this will begin the evolution of life in that planetary system. These bacteria will play the role of LUCA (Last Universal Common Ancestor) for the local evolution of life. For this to happen, the bacteria must be similar to Earth's endolithic extremophile bacteria, which can build their chemosynthesis directly on dissolving minerals and actively reproduce at temperatures close to the boiling point of water. Bacteria are likely to be accompanied by viruses that cohabit with them, so primary biospheres will contain both bacteria and viruses. An intelligent civilization with relatively modest technological development, not too far above Earth's civilization, would be sufficient to implement such a program.

Other scenarios are conceivable as well. For example, a civilization might send its robotic agents into the region of a forming or very young planetary system, where they would control the natural evolution of life by methods simi-

lar to artificial selection. Such selection might allow the 'breeding' of life with a genetic code containing the desired information patterns. Implementing such a program requires somewhat more subtle work than in the previous scenario, but again, nothing seems impossible.

Thus, the objection that the project is 'fantastic' is unlikely to work. However, beyond the information signatures of the universal genetic code *per se*, what arguments can we have for discussing their possible accidental or artificial nature? Does such a discussion have a perspective? In this regard we will discuss several separate issues, organized in the following subsections.

9.1. Meaningful Information Structures

Throughout the decades of work on the SETI problem, the question of what is the criterion for the artificiality of the signal has repeatedly arisen, and the concepts of it have gradually changed. Many curiosities have been associated with this fact. For example, when radio pulsars were discovered, their discovery was not reported for a long time because the suspicion that artificial signals from extraterrestrial intelligence were being observed was quite strong. There have even been cases where physical criteria for the artificiality of signals were proposed in advance, but then quite natural objects were discovered that met those criteria. For example, the presence of emission lines of technetium in starlight: technetium has no stable isotopes and therefore should not occur in nature. Another example is the presence in the spectrum of the source of emission lines of chemical elements with a Doppler shift of opposite sign: after all, it is not possible for an object to be approaching and moving away at the same time! Over time, however, stars with spectra containing technetium lines and sources with Doppler lines with opposite shifts have been discovered (*e.g.,* the famous microquasar SS-433).

Over time, it has been recognized that the only reliable criterion for the artificiality of a signal is the separation of meaningful information within it. In creating interstellar messages, we ourselves follow the prescription of beginning such messages with easily interpretable patterns of a mathematical nature. Not only must such meaningful patterns of information be present in the message, but they must be very easy to read and clearly declare their artificiality. For example, the interstellar messages Cosmic Call 1 and Cosmic Call 2 (Dumas 2007) began with a series of natural numbers and then a series of prime numbers. Arecibo's Cosmic Message also begins with a series of natural numbers, followed by the atomic numbers of the basic 'elements of life': hydrogen, carbon, nitrogen, oxygen, and phosphorus. This is easy to understand, and the artificiality of the pattern is hard to doubt.

Are there meaningful information structures among the information signatures found in the genetic code? Compared to the expected messages of extraterrestrial intelligence, the problem of genetic code information is that there is

no natural 'reading order', so meaningful patterns should be searched everywhere in the code. It should be noted that among the SM-signatures found, all the numbers 111, 222, 333, 444, … , 999 are found without exception. Can this be interpreted as a representation of the first nine numbers of the natural series? The question is open. Of course, they are not in the order of the natural series, but this does not cast a shadow on the set of these signatures as representatives of the natural series. Among the signatures found, there is nothing conspicuous like a series of prime numbers as in the Cosmic Call, or atomic weights as in the Aresibo Message, or, for example, a series of Fibonacci numbers. However, 'explicitly' there is a reference to the pair (10, 37), which has non-trivial mathematical properties, and there is also a structure representing the Pythagorean Theorem: $3^2 + 4^2 = 5^2$. So it is impossible to say with certainty that there are no meaningful 'texts' in the genetic code, even though the available patterns seem to be 'weaker' than, for example, the Cosmic Call series of prime numbers.

In our analysis, the 'information levels' 1, 2, 3, 4, 5 emerged naturally, in the order of increasing complexity of information and in the order of using more and more resources to present it, including the first two very simple 'attention signal' levels. A similar structure is to be expected for real cosmic messages, but what are these levels of information really in our case? Are they simply a convenient way of categorizing and listing information patterns of a particular kind, or are they really an indication that 'this was planned on purpose' and that we have unraveled the extraterrestrial intent? Again, it is not possible to claim that the used division into information levels is a fiction, but it is also not an unambiguous 'proof of artificiality'. Yes, in fact, in the order of searching and examining the signatures of the genetic code, we proceeded quite naturally through the information levels, as if they had been deliberately prepared for our analysis. However, all the 'naturalness' of this process does not prove that these levels are really artificial in origin and do not exist only in our imagination.

If in the cells of the table of the genetic code (see Fig. 1) instead of the names of amino acids we write the weights of their side chains, then, taking into account all the signatures in the form of SM-numbers and various 'equilibria of weights', the resulting numerical table will resemble the result of solving a very complex puzzle, a bit like Sudoku (but much more complicated). The paper (Shcherbak and Makukov 2013) explains that this 'puzzle' can be formulated as a problem of solving a system of linear Diophantine equations and inequalities with the additional condition that the solution is sought among a fixed set of natural numbers. This idea explains how the found signatures of the genetic code could be practically constructed if they were indeed artificial. However, even this does not prove that the whole set of signatures found is really the result of purposeful planning and makes sense.

It might seem natural for extraterrestrial intelligence messages to rely solely on the binary number system, which, like some other fundamental mathematical structures, should have universal meaning for any intelligence. Then the choice of base 10 to represent the signatures of the genetic code should be considered unnatural, and the real prominence of the positional number system with base 10 in the signatures found should speak against the presence of meaning in them and against their artificial nature. However, this strong objection is suddenly removed by the simple fact that our own cosmic messages, such as Cosmic Call 1 and 2, are based on the use of the decimal number system. The first lines of Cosmic Call messages define the decimal representation of the natural series, after which only the decimal number system is used throughout the message. And this actually makes sense. The use of the decimal number system says a lot about us: that we are not computers (which would definitely prefer the binary number system), that ten is somehow preferable to us. We probably have ten special limbs of some kind that are convenient to count – maybe it is ten tentacles, maybe it is ten fingers, maybe it is ten of something else that is convenient to count.

It can be said that the situation turned out to be more complicated than the criterion of meaningfulness as a criterion for the artificiality of SETI messages was supposed to be. The question of the presence of meaningful information structures in the case of the genetic code is not resolved unambiguously. After studying the information signatures of the genetic code, we are left in a state of agonizing uncertainty about the absence or presence of meaningful information in it. As is often the case, the reality turns out to be richer than any expectations about it. This is also an important lesson for the SETI problem in the general case: the question of meaningfulness may not be resolved unambiguously when receiving signals from space that would be considered as a candidate for a message from extraterrestrial intelligence.

9.2. On the Estimation of the Degree of 'Improbability' of the Information Signatures of the Genetic Code

We do not know any natural explanations for the whole set of the Rumer symmetries, the appearance of a large number of independent SM-numbers among partial sums over different *a prior*i selected subgroups of coding products and other 'probabilistic miracles'. Probably they simply do not exist. The question remains, what is the degree of accidental nature of the totality of the observed information patterns, if all observed information signatures are indeed in some sense accidental? May it be that the probability of the entire observed pattern is so absurdly small that its accidental origin should be simply excluded?

We have already started to estimate the degree of randomness of the signatures obtained in Sections 3 and 4, where we discussed the first two information

levels of the genetic code. From these two levels we can obtain a simple estimate for the cumulative probability of the signatures of the first two levels: $1/1024 \times 1/74 \times 1/74 \times 1/32 = 6 \times 10^{-9}$. For the third and fourth information levels (Sections 5, 7), 17 independent SM-numbers were obtained. Since an SM-number is obtained with a probability of 1/111 by randomly selecting an integer, the total probability of obtaining 17 independent SM-numbers is $(1/111)^{17} \sim 2 \times 10^{-35}$. This gives a total scale probability of 10^{-45}, and that is without considering several 'equilibria' that are also independent signatures and the probability of the Calligram-B symmetrical patterns (see Section 8). As the total probability of the entire observed pattern, we get a really absurdly small value, still several orders of magnitude smaller than 10^{-45} (probably something like 10^{-50}). The accuracy of the calculation does not play an important role here, since this value is not a realistic estimate of the degree of 'improbability' of the observed coincidences, and it is actually fundamentally impossible to obtain a 'correct' estimate of the degree of this 'improbability'.

The difficulty lies in the aforementioned 'search at corners' problem, which is well known in experimental physics. In the simplest case, it looks like this.

Suppose you measure an experimental curve $Y(X)$ and expect it to behave smoothly within the statistical errors of the measurement. But instead of the expected smooth behavior, you discover something more complicated, such as two large upward outliers, Y_1 and Y_2, at two different points on this curve, X_1 and X_2. You need to determine the probability that the observed pattern is not random in order to figure out what happened: did you discover a new effect, or is it just a statistical fluctuation? It is not difficult to find the probabilities that the amplitude of your curve will be at least as large as Y_1 and Y_2 at points X_1 and X_2, respectively. If you multiply the probabilities, you will find the total probability that both jumps will occur at the same time. You will get a small probability (if the jumps are really large), but it will not be the correct answer to the question of how unlikely the observed pattern is, if it is indeed random. The point is that random jumps can occur at other values of parameter X, and you need to determine the probability of a pattern similar to your result occurring with any combination of parameters X_1 and X_2. This is already much harder to do, simply because it is sometimes quite difficult to give a precise definition of what a 'similar result' means. The probability for such an eventuality, taking into account all similar results, will be much larger than the first naive estimate. In physics, in such cases, you have to find some workarounds to get probability estimates (the most common one is to use chi-square criterion or something like that), but they can usually be found (not always easily). That is, you need to estimate not only the probability of the exact result you have, but also the joint probability of getting any similar result. This is what 'search at corners' is all about.

In computer science or linguistics, a similar problem can arise, but it is much harder. Let us take a primitive example. Suppose we put a monkey in front of a computer with a word processor and let him pound the keys until he fills a whole page. We examine the result of his work and suddenly, to our amazement, we find that an entire line of characters (80) represents some meaningful text. Question: How improbable is an event we observe? Obviously, it is meaningless to estimate the probability of generating exactly the text that resulted, even taking into account the fact that the same meaningful piece of text could have resulted in different places on the page. The correct estimate would be to estimate the probability that *any* meaningful line of text at least 80 characters long will randomly appear anywhere in the text. But we have no way to enumerate all the meaningful texts to count them and find the probability we need. The problem is that we do not have a common definition of meaningful text: for example, how many grammatical errors are allowed for a text to count as meaningful, what exactly is meaningfulness, and so on. Is the string 'Twas brilling, and the slithy toves did gyre and gimble in the wabe' a meaningful text? So our problem in this case is fundamentally intractable. The problem is not that the problem is difficult, but that we cannot even formulate the problem.

Similarly, it makes no sense to estimate the probability of occurrence of the exact set of information signatures observed in the universal genetic code, since we can imagine many other signatures that we would also subjectively consider unlikely and strange. And they may be arranged in a different way than the one we really observed. To estimate probability, we should use them all, and consider what we actually have as only one representative of this huge class of objects. The correct description of probability should be that we go through all the codes that give signatures at least as strong as the one we found, and divide the number of such codes by the number of all probable genetic codes (still satisfying some criterion of sufficient optimality, which is also not too easy to formulate). But it is impossible to determine exactly what 'signatures of no less strength than the detected one' means because of their almost infinite variety, so there is no way to count the number of genetic codes that favor such signatures.

The conclusion is that we have no objective way to assess the degree of 'improbability' of the 'high regularity' of the genetic code that we have observed, and therefore any such assessment remains irreducibly subjective, and cannot help us choose between the alternatives of whether the information structures we have discovered are of accidental or artificial origin.

9.3. What to Do? Is There Any Prospect of Solving the Accidental/Artificial Problem?

After reading Sections 9.1. and 9.2., there may be a sense of an epistemological impasse in resolving the accidental/artificial dilemma, but this would not be entirely true, since there are some ways to change the *status quo*.

The hypothesis of the artificiality of the origin of the information signatures of the genetic code has verifiable predictions that, in principle, can be directly compared with observations and, possibly, refuted by them. If this happens, the hypothesis of the artificiality of signatures will also be disproved (Popper's falsifiability principle is used here). This way of testing the artificiality hypothesis has been proposed in the work by Makukov and Shcherbak (2018).

At present many alternative rare genetic codes are known.[7] It is assumed that all alternative genetic codes are late mutations of the main universal genetic code (Frank-Kamenetsky 2018). This assumption is shared by most scientists. If information signatures were artificially introduced into the basic universal genetic code, they were introduced only once, so all subsequent mutations of the genetic code could only corrupt these information signatures, but could create something essentially new only with a very low probability. This prediction is directly verifiable.

We have examined all 25 alternative genetic codes available to us. In eight of them, the basic Rumer structure of two equal octets was broken, so the preservation of the other signatures cannot even be questioned. All remaining alternative codes destroy a very important signature of the second information level – simultaneous divisibility of the masses of octets **I** and **II** by 74, not to mention simultaneous divisibility by 740 and an indication of the special role of the pair (10, 37) in the whole problem, if the value 74 is used for the total weight of the **stop** signal. This is also true for the *euplotidium* genetic code, which was discussed in Section 5 in relation to the TGA switch. In this sense, the *euplotidium* code is no more special than any other genetic code.

This confirms the prediction of the artificiality hypothesis that the information signatures of the universal code are corrupted in alternative codes. However, it is still possible to search for some completely new information signatures in alternative codes that are in no way related to the signatures of the universal genetic code. We have not tried to solve this problem, because it is even difficult to formulate it: strictly speaking, it is necessary to find unknown what. Nevertheless, the possibility of falsifying the predictions of the hypothesis of the artificiality of the signatures of the universal code remains. The predictions will be falsified if someone finds and presents alternative information signatures in an alternative code that are comparable in strength to the information signatures found in the universal genetic code. Here, however, the question of the degree of objectivity of such a comparison will arise, but we will not touch it.

It is important that the proposed way gives the possibility of experimental verification of the hypothesis of artificiality of signatures, therefore this hy-

[7] Twenty-five variants of genetic code modification can be found at URL: https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi.

pothesis has the status of a scientific statement in the sense of Popper's principle of falsifiability. The hypothesis of the artificiality of signatures of the genetic code withstands the verification by observations so far, so it deserves the right to exist together with the alternative assumption that the whole collection of discovered signatures is nothing more than an accidental. As new modifications of genetic codes are discovered, and as alternative codes are studied more deeply for the presence of information signatures, the hypothesis of the artificiality of information signatures of the universal genetic code will be subjected to new verification. One day the hypothesis may be disproved, but until then its plausibility will only increase.

To be fair, it should be added that the accidental hypothesis of genetic code signatures, unlike the artificial hypothesis, is not falsifiable in the Popper's sense, even though it may seem more natural and viable to many. The accidental hypothesis does not lead to predictions that can be meaningfully falsified. In response to the presentation of any convincing 'meaningful' information signature, there is always the freedom to reply that it is nothing more than the result of chance. In this sense, the accidental hypothesis is not a strictly scientific hypothesis, unlike the hypothesis of artificiality, which in no way casts a shadow on the hypothesis of chance. Not all statements that cannot be strictly classified as scientific hypotheses are in any way flawed or useless. An example is Darwin's hypothesis of natural selection, which is also unfalsifiable.

Another possibility for the development of the problems discussed in this article is to continue the search for information patterns in the universal genetic code and in the molecular machinery directly related to the genetic code. We are far from being sure that all interesting information signatures have already been identified. First of all, one should pay attention to the very complex structures of tRNA and ARS, which are directly related to the genetic code, as explained above (see Section 8). With regard to the search for strange information patterns, these structures are poorly studied, and who knows, surprises may await us here. There is no absolute certainty that the problem has been studied exhaustively and at higher levels of information, so work can continue in these directions as well.

10. Conclusion

In Sections 3–8 we have given a detailed overview of the strange symmetries and information patterns that can be found in the genetic code and in molecular structures directly related to the implementation of the genetic code. It turns out that the genetic code contains surprisingly many such signatures. Moreover, we found that these symmetries and information patterns are naturally distributed in five information levels in order of increasing complexity of information representation and in order of using new resources for its representation. Furthermore, the first two levels are very simple and natural and look

like ‘an attention signal’ whose presence is usually assumed in messages from extraterrestrial intelligence.

These symmetries and information signatures cannot be understood in terms of the functionality of the genetic code, so their appearance in the genetic code can be explained either by mere chance or by their artificial origin.

The most reliable criterion of artificiality would be to identify the semantic component in information signatures, but the situation turns out to be confusing. We cannot insist with certainty that there is no semantic content in the signatures, because some simple ‘semantic patterns’ are definitely can be detected (see Section 9.1.).

To assess the realism of the artificiality hypothesis, we cannot rely on the small probability of accidental appearance of information signatures of the genetic code, because it is impossible to quantify this probability, since the problem of estimating this probability cannot be formulated correctly (see Section 9.2.).

Some perspective for solving the accident/artifact dilemma is provided by the fact that the artificiality hypothesis makes a prediction that can be directly verified by observations. The essence of this prediction is that no rare alternative genetic codes will, first, contain all the information signatures of the universal genetic code and, second, contain no strong alternative information patterns. Currently, 25 known alternative genetic codes have been tested for the first half of this prediction, but not for the presence of alternative information structures. The first half of the prediction is confirmed (see Section 9.3.). The second half of the prediction remains to be verified. There exists also the possibility of looking for new, deeper information signatures in the basic universal genetic code. This provides a possible direction for the development of the issues considered in this paper.

Acknowledgement

The authors would like to thank Maxim Makukov for an extremely fruitful discussion.

References

Crick F. 1968. The Origin of the Genetic Code. *Journal of Molecular Biology* 38: 367–379.

Danckwerts H. J., and Neubert D. 1975. Symmetries of Genetic Code-Doublets. *Journal of Molecular Evolution* 5: 327–332.

Dumas S. 2007. *The 1999 and 2003 Messages Explained*. URL: https://www.plover.com/misc/Dumas-Dutil/messages.pdf.

Frank-Kamenetsky M. 2018. *The Most Important Molecule: From DNA Structure to XXIst Century Biomedicine.* Alpina Publisher.

Hoesl M., Oehm S., Durkin P., Darmon E., Peil L., Aerni H., *et al.* 2015. Chemical Evolution of a Bacterial Proteome. *Angewandte Chemie* 54: 10030–10034.

Hohsaka T., Ashizuka Y., Murakami H., and Sisido M. 2001. Five-Base Codons for Incorporation of Nonnatural Amino Acids into Proteins. *Nucleic Acids Research* 29: 3646–3651.

Konopelchenko B., and Rumer Yu. 1975. Classification of Codons in the Genetic Code. *DAN SSSR* 223: 471–474. *In Russian* (Конопельченко Б., Румер Ю. Б. Классификация кодонов в генетическом коде. *ДАН СССР* 223: 471–474).

Makukov M., and Shcherbak V. 2018. SETI In Vivo: Testing the We-Are-Them Hypothesis. *International Journal of Astrobiology* 17: 1–20.

Marx G. 1979. Message through Time. *Acta Astronautica* 6: 221–225.

McMurry J. 2023. *Organic Chemistry: A Tenth Edition.* OpenStax, Rice University, Houston, Texas.

Nikitin M. 2016. *Origin of Life. From Nebula to Cell.* Moscow: Alpina non-fiction. *In Russian* (Никитин М. *Происхождение жизни. От туманности до клетки.* Москва: Альпина нон-фикшн).

Pacioli L. 1508. *De Viribus Quantitatis.* Library of the University of Bologna.

Rumer Yu. 2013. Systematization of Codons in the Genetic Code. *DAN SSSR* 183: 222–226. *In Russian* (Румер Ю. Систематизация кодонов в генетическом коде. *ДАН СССР* 183: 222–226).

Shcherbak V., and Makukov M. 2013. The 'Wow! Signal' of the Terrestrial Genetic Code. *Icarus* 224: 228–242.

Wilhelm T., and Nikolajewa S. N. 2004. A New Classification Scheme of the Genetic Code. *Journal of Molecular Evolution* 59: 598–605.

Zaitsev A. 2008. The First Musical Interstellar Radio Message. *Journal of Communications Technology and Electronics* 53: 1107–1113.