Scientific Research Publishing

Journal of Computer and Communications, 2018, 6, 64-83
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

The Study of the Secrets of the Genetic Code

N. N. Kozlov

Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia
Email: gencodkiam@mail.ru

How to cite this paper: Kozlov, N.N. (2018) The Study of the Secrets of the Genetic Code. *Journal of Computer and Communications*, **6**, 64-83.
https://doi.org/10.4236/jcc.2018.67007

Received: June 26, 2018
Accepted: July 28, 2018
Published: July 31, 2018

Abstract

The disclosure of many secrets of the genetic code was facilitated by the fact that it was carried out on the basis of mathematical analysis of experimental data: the diversity of genes, their structures and genetic codes. New properties of the genetic code are presented and its most important integral characteristics are established. Two groups of such characteristics were distinguished. The first group refers to the integral characteristics for the areas of DNA, where genes are broken down in pairs and all 5 cases of overlap, allowed by the structure of DNA, were investigated. The second group of characteristics refers to the most extended areas of DNA in which there is no genetic overlap. The interrelation of the established integral characteristics in these groups is shown. As a result, a number of previously unknown effects were discovered. It was possible to establish two functions in which all the over-understood codons in mitochondrial genetic codes (human and other organizations) participate, as well as a significant difference in the integral characteristics of such codes compared to the standard code. Other properties of the structure of the genetic code following from the obtained results are also established. The obtained results allowed us to set and solve one of the new breakthrough problems—the calculation of the genetic code. The full version of the solution to this problem was published in this journal in August 2017.

Keywords

Genetic Code, Overlapping Genes, Degeneracy Code, Code Irregularities, Potential Code for Overlaps, Common Property Codes, Integral Characteristics Code, Deviation Code, Elementary Overlaps, Role-Reinterpreted Codons, Code Is Not Arbitrary

1. Introduction

According to the results of the research started in 1992, the author developed a

mathematical theory of genetic code, briefly presented in the article N. N. Kozlov and T. M. Eneev, The Fundamentals of a Mathematical Theory of Genetic Code. Doklady Mathematics 2017, Volume 95, № 2, pp. 144-146. This work is an extended version of this article. The focus will be on the issues that have received structured mathematical justification. We tried to give the material in a form accessible to a wide range of mathematicians, so the first two introductory paragraphs are given. The task is simplified by the fact that the extended versions of some of the main results were published in a number of articles cited in the list of references. However, it is crucial to present such results in their relationship, because they relate to the mysteries of one and the same mysterious structure—the genetic code. It should be noted that none of these results was found by the author in other publications, both in domestic and in foreign ones. The disclosure of many secrets of the genetic code was facilitated by the fact that it was carried out on the basis of mathematical analysis of experimental data: the diversity of genes, their structures and genetic codes identified by a number of features.

2. Genetic Code

The history of the discovery of the genetic code is described in detail By M. Ichas [1] [2]—one of the participants of the pioneer research on this problem. He writes: "… decoding of the biological code is a revolutionary event, it may be appropriate to compare it with another event that caused a revolution in science a hundred years ago with the appearance of Darwinian 'Origin of species' [1]. The most difficult part of the code problem was to understand that the code exists. It took almost a century. Its counting is conducted from the work of Mendel [3], which showed that hereditary characteristics are transmitted by discrete parts, which we today call genes. This work, as we know, almost did not arouse interest. From what we know, it seems that Mendel was, in General, indifferent to the responses to his work. Having published his main work, he considered his duty to be complete: if it did not pay attention, the worse for readers, and not for the author." ([2], p. 142). In 1900, three independent researchers simultaneously confirmed the results obtained by Mendel with their experiments. Only after completing the work, they learned that 34 years ago they were ahead of Mendel. After 1900, genetics began to develop rapidly and continuously.

For the first time the idea of molecular biological approach to the problems of genetics was formulated by the famous physicist E. Schrödinger in his book "what is life? From the point of view of physics" [4], which in original saw the light in 1945. On page 28 read the code overview (21 years before the final solution!): "In calling the structure of the chromosomal threads cryptographic code, we mean that sekwati reducing mind like that once imagined by Laplace and to which every causal connection directly open, could is—the pacing of the structure of chromosomes, say, will develop if the egg under favorable conditions, into a black cock or a speckled hen, into a fly or a maize plant, a rhododendron, a beetle, a mouse or a man." In addition to this, and other brilliant predictions, it

should be noted that this book has played a crucial role in the fate of a number of theoretical physicists. I will name only two names, which will be discussed in the future. This is F. Creek, who in 1946 left theoretical physics and turned to the problems of biology after reading this book. His Nobel lecture was devoted to the problem of code, not the structure of DNA, for which he was awarded the Nobel prize (F. Crick—Nobel Lecture, Dec. 11, 1962: on the Genetic Code, Internet). The origin of the code problem was also the physicist G. Gamow, Which F. Creek refers to on the first page of the lecture. However, at the final stage of research, it was the biochemists who experimentally established the genetic code and indicated its role in protein biosynthesis. For this result H. Koran, M. Nirenberg and R. Holly received the Nobel prize in 1968.

But first the problem of DNA structure was solved. It took about a year and a half for a graduate student D. Watson to solve one of the most important problems of biology, which is now considered one of the main fundamental problems solved in the last century, together with the Director F. Crick. We are talking about the structure of molecules of DNA, which the world saw for the first time on 25 April 1953: [5], a one (!) the "Nature" journal page put an end to the discussion on the role of DNA in the transmission of hereditary information. Descriptions that give DNA today is different. For our purposes, a simplified description is enough. The model of the DNA double helix is two strands twisted relative to each other (Figure 1).

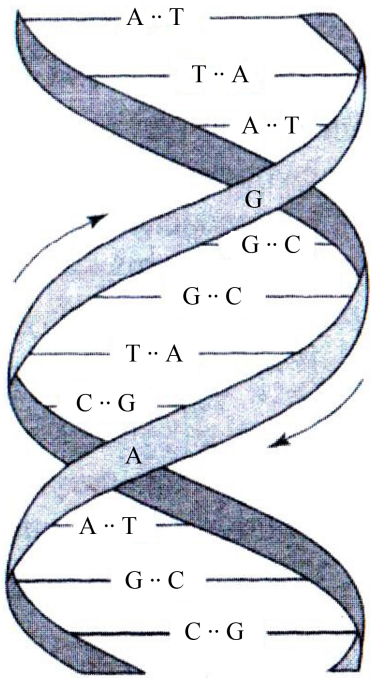

Figure 1. Model of DNA double helix. The direction of the reading if the gene text is indicated by arrows, from top to bottom in one strand and from bottom to top in the other.

In fact, it is a double helix, not any spiral. DNA contains only 4 letters: A, C, G, T. These are four nucleotides: adenine, cytosine, guanine, and thymine, respectively. The points between these letters in **Figure 1** indicate the number of hydrogen bonds: two bonds between a and T and three between C and G. it is this brilliant conjecture of Watson, who introduced these complementary pairs [6], and allowed to explain the important properties of the transmission of hereditary information. (These connections exist between the two DNA spirals.) DNA is measured in different ways, including the number of nucleotide pairs. For example, for human DNA there are about 3.2 billion (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/).

The secret of the gene was finally solved in 1966 (by the centenary of Mendel's work [3]), when in the course of experimental studies it was finally established that the genes are single-stranded parts of DNA and contain information about the protein in coded form. It turned out that each of the 20 amino acids—the elements that make up all the known proteins—is encoded by certain triples of nucleotides-codons or triplets. For four letters: A, C, G, T, we have 64 codons: AAA, AAC, AAG, …, TTT. The meaning of all these codons was experimentally established and presented in the table of the genetic code, and the encoding, which nature chose, turned out to be peculiar (**Table 1**). It is presented completely.

Table 1. Standard genetic code.

	1	2	3
1	Met	1	ATG
2	Trp	1	TGG
3	Phe	2	TTY
4	Tyr	2	TAY
5	His	2	CAY
6	Asn	2	AAY
7	Asp	2	GAY
8	Cys	2	TGY
9	Gln	2	CAX
10	Lys	2	AAX
11	Glu	2	GAX
12	Ile	3	ATM
13	Val	4	GTN
14	Pro	4	CCN
15	Thr	4	ACN
16	Ala	4	GCN
17	Gly	4	GGN
18	Ser	6	TCN, AGY
19	Leu	6	CTN, TTX
20	Arg	6	CGN, AGX
	ter(*)	3	TAX, TGA

Note. For each of the amino acids are given: 1—standard three-letter abbreviations, 2—the number of codons-synonyms, 3—three-letter nucleotide representations of codons. Designation: X: A, G; Y: T, C; M: T, C, A; N: A, G, T, C. The last line contains three terminator codons—ter, each of which denotes the stoppage of protein synthesis.

It turned out that only two amino acids—methionine (Met) and tryptophan (Trp) are uniquely encoded by ATG and TGG codons, respectively. All other amino acids are encoded by more than one codon (these are co-dons-synonyms), but not more than six. The latter is observed only for three amino acids: serine (Ser), leucine (Leu), arginine (Arg). Such three encodings are called irregular, unlike 17 others, the regulars for which every 1st and 2nd positions are the same in the corresponding set of codons-synonyms. The total number of semantic triples, or codons that encode any amino acid, as well 61. Ter (*) term-nation codons do not correspond to any amino acids, each of them stops protein synthesis. In view of the importance of these codons in further analysis, we select them in (1):

$$\text{ter: TAA, TAG, TGA.} \quad (1)$$

We point out that in addition to degeneracy (*i.e.*, when several co-dons-synonyms correspond to the same amino acid), the most important property of the code is its universality: the code is the same for almost all living organisms. However, to date, a number of deviations from the standard code have been found, which is one of the most mysterious features of the code (see Table 2).

Table 2. Standard K^0 and non-standard genetic codes K^1 - K^{14}. and their integral characteristics—*p*.

1	2	3	*p*
K^0	The standard code		16
K^1	The Vertebrate Mitochondrial Code	TGA(ter) → Trp, ATA(Ile) → Met, AGX(Arg) → ter	7
K^2	The Invertebrate Mitochondrial Code	TGA(ter) → Trp, ATA(Ile) → Met, AGX(Arg) → Ser	7
K^3	The Echinoderm and Flatworn Mitochondrial Code	TGA(ter) → Trp, AAA(Lys) → Asn, AGX(Arg) → Ser	5
K^4	The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code	TGA(ter) → Trp	6
K^5	The Ciliate, Dasycladacean and Hexamita Nuclear Code	TAX(ter) → Gln	5
K^6	The Euplotid Nuclear Code	TGA(ter) → Cys	5
K^7	The Alternative Yeast Nuclear Code	CTG(Leu) → Ser	16
K^8	The Ascidian Mitochondrial Code	TGA(ter) → Trp, ATA(Ile) → Met, AGX(Arg) → Gly	7
K^9	The Alternative Flatworn Mitochondrial Code	TGA(ter) → Trp, AAA(Lys) → Asn, TAA(ter) → Tyr, AGX(Arg) → Ser	0
K^{10}	Blepharisma Nuclear Code	TAG(ter) → Gln	10
K^{11}	Chlorophycean Mitochondrial Code	TAG(ter) → Leu	10
K^{12}	Trematode Mitochondrial Code	TGA(ter) → Trp, AAA(Lys) → Asn, ATA(Ile) → Met, AGX(Arg) → Ser	6
K^{13}	Scenedesmus Obliquus Mitochondrial Code	TAG(ter) → Leu, TCx(Ser) → ter	10
K^{14}	Thraustochytrium Mitochondrial Code	TTA(Leu) → ter	21

Note. 1—codes K^0 - K^{14}, 2—their names, 3—deviations from the standard code Columns 2 and 3 were obtained on the basis of: http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t.

3. Overlapping Genes

Mathematical analysis of the structure of the genetic code, first of all, based on the study of some unusual ways of recording genes encoding proteins. An unusual way to record genes is that the same area of the DNA chain encoding the protein can be shifted with the phase shift by +1 or −1 nucleotide or with another DNA chain (with or without such shifts). In other words, the same DNA sample can encode two or more proteins, up to six. Such genes have been called overlapping. Note that experiments show that such reading is allowed only in some cases, and in the vast majority of cases there is a ban on these alternative readings. This prohibition is that these shifts lead to completely different sequences of codons than the original sequence (when there are no shifts). But it was found that in such alternative sequences any codons from (1) necessarily arise; this is how the gene encoding protein is arranged, or so the corresponding codes of Amy-NOx acids are chosen due to the degeneracy of the code. Because of the role of these three codons (they stop (block) protein synthesis), protein is not synthesized in alternative reading. The conclusion was made about the powerful biological protection: nature does not need ephemeral proteins, it does not synthesize proteins corresponding to the shifted positions (for example, if the initial point moves during mutations). Thus, two reading frames—RS were introduced: an open reading frame (ORS)—a sequence of codons that does not contain codons of the term and a blocked RS-BRS, when such codons occur [7]. For rice 2 protein corresponds to ORS, shifted positions of nucleotide investigations, both +1 and −1-BRS.

In the shifted States, we obtain other sequences of codes (two other RS), in each of which the codons of ter termination will be repeatedly present, which are shown in **Figure 2** marked with the same symbol*. It will be two RS with locks-BRS. It can be seen that with the help of three nucleotide substitutions for the same Nucleo-tid C (the positions of the substitutions are indicated below the gene text), none of the three codons ter (symbol*) on a given section of the gene will not occur, and with such substitutions the protein sequence will not change because these three men represent the three substitutions of codons to synonyms. However, ti-typical gene arranged so that these shifts were given just two BRS [7]. It turned out that only for overlapping genes such a ban does not exist. For the first time this effect was experimentally installed in 1976 in the course of research in reading first whole genome of bacterial viruses FKH 174 [8]. After these studies, their leader F. Sanger became the only two-time winner in the history of the Nobel prize in chemistry. F. Sanger showed interest in one of the first of our RA-bot in a letter to him I formulated a new property of the first whole of genome [9]. It consists in the fact that to record such a genome, it is necessary to use all 61 semantic codons—this is due to the overlap of genes first discovered in this genome. I was asked to present the result in NAURE. His answer is given in the Sanger file. The total nucleotide sequence of ring single-chain DNA (APPARENTLY, the single-chain factor of DNA established earlier experimentally

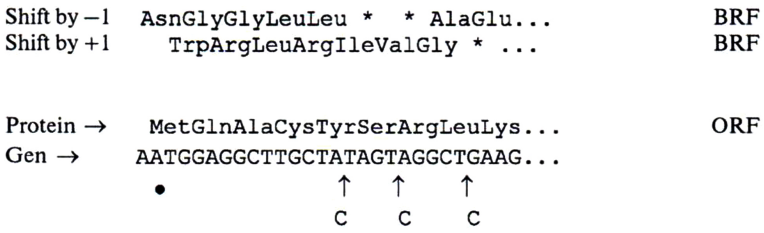

Figure 2. The plot of a protein sequence (the first amino acid is Met) encoded in the gene starting with ATG (the first nucleotide and in this triplet marked with a fat point) and cases of shifts to +1 or −1 nucleotide.

and was decisive for reading the first whole genome) contains 5386 nucleotides [10], but the total number of amino acid residues in the aftereffects of all proteins multiplied by 3 (taking into account the non-coding regions) exceeds this number of nucleotides. It has been shown that gene E contains a 273 nucleotide and the gene is localized within the D [8]. This is the first experimentally detected overlap shown in **Figure 3**.

Currently, it is believed that overlapping genes are although unusual, but still quite common element of the genome organization. In decoded the human genome discovered multiple genetic overlap [11] have been about 1700. Accumulated extensive material on the genetic overlap has set itself the goal of a thorough and comprehensive analysis. Let us focus on some of the results obtained by us on the basis of mathematical analysis.

It can be seen that there are only 5 different cases of overlapping genes, resolved by the structure of DNA (**Figure 4**), of which the first two relate to overlaps of genes from the same DNA chain, and the remaining 3-to overlaps of genes taken from different DNA chains.

For **Figure 4**, only small fragments of real interruptions are presented, and the total length of some of them reaches almost 1300 nucleotides. In addition, the total length of overlap can reach more than half of the genome size (GSHV virus).

It should be emphasized that it is the analysis of multiple relationships of co-dons in genetic overlaps that is the main tool of the conducted research.

4. Degeneracy of Code

One of the tasks that was set by us, refers to the fundamental problem of the genetic code: why do we need the degeneracy of the code, when for the same amino acid are usually more than one encoding, up to 6-and coding. We have investigated the participation of all of the sense codons in overlapping genes. It turned out that there are many overlapping genomes in which all 61 semantic codons must necessarily participate, and with the exclusion of at least one codon, the record of genetic overlap found in experiments seems impossible. One of these genomes is the first whole genome for the bacterium virus, PH174, containing overlaps for 814 nucleotides [10]. Our article [9], as well as the article accompanying it, were cited at least 100 times each (see SITA-tion file 300).

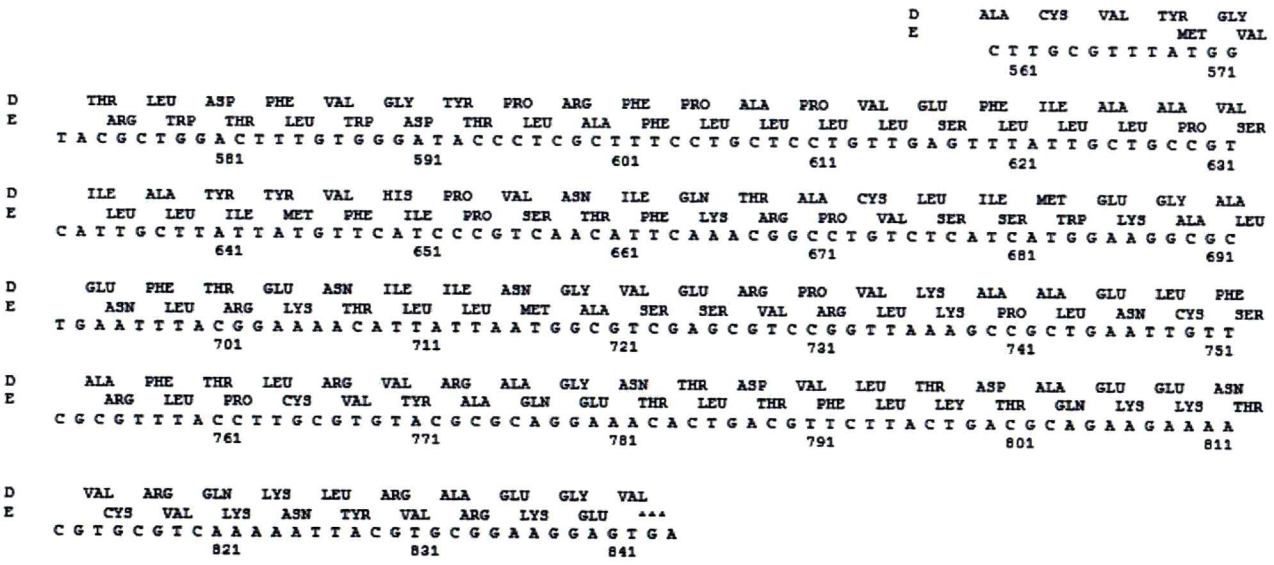

Figure 3. The first genetic overlap was found experimentally [8]. The figure is presented in the format of the publication of the full text of the first whole genome for the bacteriophage ΦX 174 [10]. We see that starting from position 568 and up to position 840, the coding of the new protein E was established on the site of the nucleotide sequence of another protein-protein D.

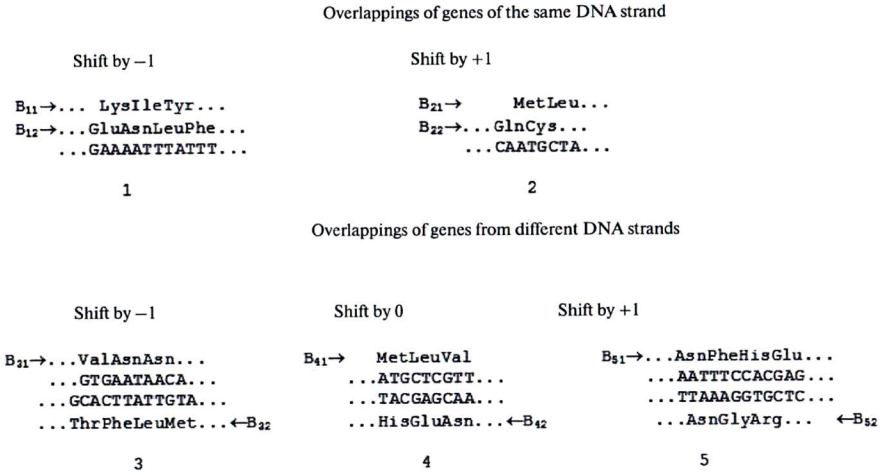

Figure 4. Five possible cases of overlaps of genes associated with a single (1,2) or two chains of DNA (3-5). Reading texts in this case is carried out in different directions (indicated by the arrow): from left to right for B11, B12, B21, B22, B31, B41, B51 and from right to left for B32, B42, B52. In these fragments contains only the canonical pair of DNA: CG and AT.

5. Irregular Code

Our next task was related to the analysis of codons that deviate the genetic code from the homogeneous structure. This is one of the most mysterious features of the genetic code. As shown by the mathematical analysis of additional codon representation from the Table 1, or irregularity, for Ser it is AGY, for Leu it is TTX, for Arg it is AGX, allow in principle "to organize" overlaps in a number of genomes or significantly expand the range of genetic overlaps both for double (for genome FX 174 in 7 times) and for triple (for HIV-2 [12] in 5 times) over-

laps, if they were organized using a homogeneous code, or code without irregularities.

6. The Potential of the Code for Many Genetic Overlaps

Next, we set the task of what is the potential of the genetic code to create all these cases of overlap. The answer was the following—a phenomenal potential! This result was evaluated by the Nobel Laureate премии. de Duve to me (see the file). It turned out that only 16 amino acid pairs out of a possible 400 can create obstacles to the construction of all 5 cases of overlap. These are amino acid pairs for only three cases of overlap:
in case 2, it is 5 pairs:

$$\text{MetMet, MetAsn, MetLys, Methyl, MetThr,} \tag{2}$$

in case of 3 it is 6 pairs:

$$\text{PheTyr, TyrTyr, HisTyr, AsnTyr, AspTyr, CysTyr,} \tag{3}$$

in case of 5, it is 5 pairs:

$$\text{PheMet, PheAsn, PheLys, PheIle, PheThr.} \tag{4}$$

In other words, it seems that the genetic code is under overlapping. Is that so? The answer to this question will be given below.

So, we have established the first integral characteristic of the genetic code, which is denoted by p and which is equal to 16 for the standard code:

$$p = 16 \tag{5}$$

7. Common Property of All Natural Codes

The result led to a halt of new tasks. What is the value of p for non-standard (deviant) codes, the number of which is 14 and continues to grow? Note that the first non-standard code was discovered in 1979 in a human cell in a separate organelle—in mitochon-drii: genes of mitochondrial DNA-mtDNA were recorded by such code [13]. Only 4 codons were reinterpreted. Calculations have shown that the value of p for all 14 deviant codes does not exceed the value of 22 or about 5% of the total number of amino acid pairs [14], see **Table 2**. At the same time over-pretime were all the same three cases of overlapping, like for the standard code and in addition was discovered a code with a zero value of the value p. Thus all the natural genetic codes have a small number of prohibitions on the construction of a genetic overlap. This may be seen as a common property of all natural codes known to date. The question arises: why do natural codes correspond to this, while the number of records of genes with overlaps is immeasurably less than the usual non-overlapping genes and what is the role of rethought codons? Both issues were resolved.

8. One Mathematical Analogy

In solving the first of these issues, an important mathematical analogy was established between gene overlaps from different DNA chains and the most impor-

tant structural units. As is known, the gene sityva is with DNA, occurs in his text multiple modification of nucleotide T with U (uracil) and forms mRNA, which in turn is structured. The most important elements of this structure are the stems-fragments containing bonds similar to those in DNA. For rice 5 one of the stems of the most known secondary structure of matrix RNA-RNA MS2 [15] is given, which contains more than 130 stems. The structure was completely analyzed by us, and the results are presented in my monograph 2014.

For **Figure 5** shows a single stalk of this secondary structure (B), there are also a fragment of A, in the range 3022-3048 primary structure in the style of the quoted article. However, in fragment B, given the record of not only the nucleotides and the corresponding amino acids.-A. But the stem of **Figure 5(b)** corresponds to the overlap of fragments as if taken from different DNA chains. The reading direction on the stalk B (arrow-Ki) becomes different and this fragment of the secondary structure of the equivalent overlap shaded in a areas taken from different DNA targets (case of overlap 4). However, there are no different chains. This effect is the rotation of the original reading direction (→) (bottom line **Figure 5(b)** (←)) is due to the presence of a so-called pin loop UCUAUAA-sequence, not the stem.

What is the role of the smallness of the first characteristic *p*? The fact that its small value allows you to build a phenomenal variety of genetic overlap genes, and thus it allows you to build a phenomenal variety of secondary structures, including also functionally significant areas of the secondary structures of mRNA for all genes recorded as standard, and any of the known deviant codes. The increase of the specified characteristic for the code deviated from the standard one (for example, by an order of magnitude, as it was shown for the hypothetical code from the monograph) leads to a significant reduction of such diversity.

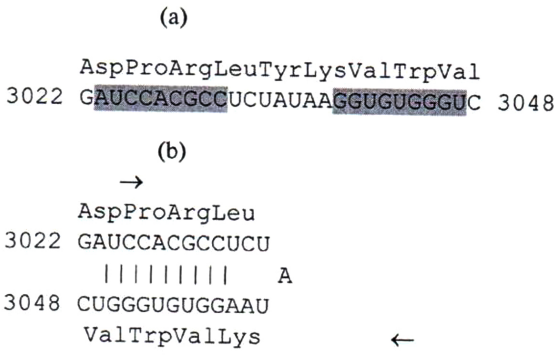

Figure 5. The stem of the secondary structure of MS2 RNA matrix. At the top (a) is given a linear text—a fragment in the range 3022-3048. And the shaded areas correspond to the stems of the secondary structure. Under the text (b), a fragment of the secondary structure is shown. Shows the presence of noncanonicial of a pair of GU (they were discovered experimentally in the structures of the RNA), in addition to the canonical CG in DNA, the canonical and analogue-AU.

Thus, it is established that the small value performs two functions: it allows to build both a phenomenal set of genetic transformations and a phenomenal variety of secondary structures of matrix RNAS for all genes.

9. About the Role Rethought Codons

Let us now consider the role of the reinterpreted codons. We raised the question of the possible relationship between the limitations on the overlap (2)-(4) and the code variability observed in a number of organizations. The analysis showed that such a relationship exists, and it is expressed in the fact that for a number of deviant codes (examples for some of them found in mitochondrial DNA are shown in Figure 6), at-home rethinking codons lead to the possibility of constructing GE-neticesi ceilings no standard code.

In each of the four pieces of overlap shown in Figure 3 the role of the same permutation is shown: TGA(ter) Trp. This natural permutation is observed for

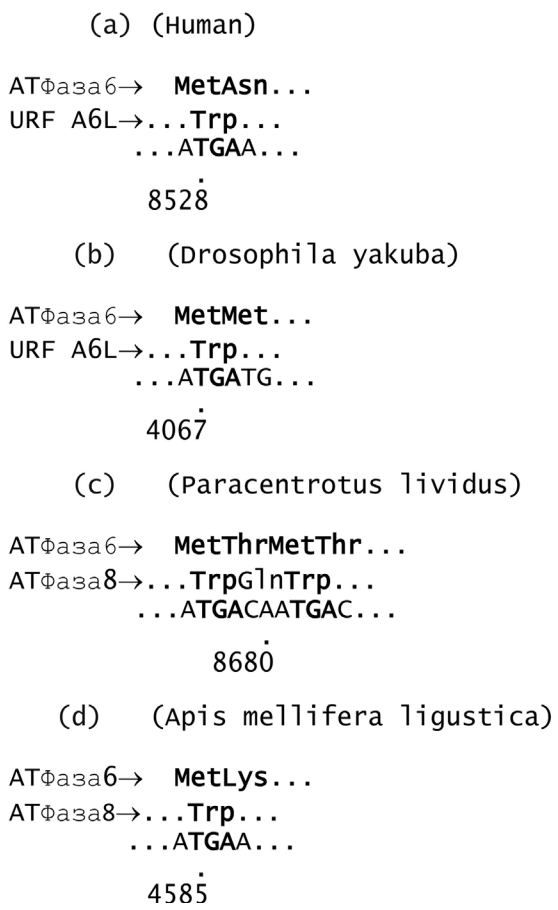

Figure 6. Fragments of genetic overlaps found in the mitochondria of four organisms whose genes are recorded by codes deviated from the standard code. This is the ceiling in one of the DNA chain. Fragments and names of proteins are given by publications [16] [17] [18] [19]. The number indicates the nucleotide number in the genome.

three deviant codes, which correspond to the given fragments, respectively; the second and fourth fragments are written by the same deviant code. Moreover, a unique permutation is present in all three deviant codes. It turned out that such a permutation made possible overlaps for MetAsn pairs (Figure 6(a), this case corresponds to the DNA of the human mitochondria), MetMet (Figure 6(b), twice MetThr (Figure 6(c) and MetLys (Figure 6(d)), which are forbidden for standard code, see (2). Specified nucleotide pairs and reinterpreted codons are highlighted. Thus, the size of genomes is reduced due to the possibility of building gene overlaps, which are not possible for the standard code. Such a reduction for a living cell can be quite large, because the number of mitochondria, as a rule, more than 1 and can reach a million. The study was cited at least 100 times (see file 300 citation).

The results obtained allowed us to turn to the analysis of experimental data on all deviant genetic codes, or codes rejected from the standard code. However, within the framework of genetic transformations, I was not able to explain the functional significance of all over-understood codons in all deviant codes. The required solution would be found in the study of areas of DNA where genes do not overlap, and such genes—the vast majority.

10. Two Integral Characteristics of the Code

We are talking about the natural blocking of the genes when all 5 codons of sequences, alternative sequences of a gene whether the reading frames—RS contain multiple stopping protein synthesis, or codons from the set (1). For Figure 7 is shown for a portion of a gene and early simplified Figure 2.

The potential of such blocking for a standard genetic code was established: it was shown that for such code only 210 amino acid pairs out of 400 possible ones participate in the blocking process. It is shown that the 31 pairs of them gives the

```
ORF0        MetSerIleLysLeuSerTyrArgGluSerPheSerIleLeuGluGluVal…

BRF1 (-1)  TyrGluHis * Thr * Leu * ArgValIle * TyrIleArgGlyGly…

BRF2 (+1)     * AlaLeuAsnLeuValIleGluSerHisLeuValTyr * ArgArgPhe…

→          TATGAGCATTAAACTTAGTTATAGAGAGTCATTTAGTATATTAGAGGAGGTTTA…
←           ATACTCGTAATTTGAATCAATATCTCTCAGTAAATCATATAATCTCCTCCAAAT…

BRF3 (-1)  IleLeuMetLeuSerLeu * LeuSerAspAsnLeuIleAsnSerSerThr …

BRF4 (0)    HisAlaAsnPheLysThrIleSerLeu * LysThrTyr * LeuLeuAsn…

BRF5 (+1)    SerCys * Val * AsnTyrLeuThrMet * TyrIleLeuProProLys…
```

Figure 7. Six RF for the gene fragment (beginning with the ATG (Met) codon, the reading direction is indicated by an arrow → one of which is open-ORFO (it has 17 semantic codons), and 5 RF—alternative RF are blocked: BRF1-BRF5. While BRF3-BRF5 correspond to the other chain of DNA and reading the sequence of codons is performed in the reverse direction (←). Figures in brackets show the shift in nucleotides relative to the ORFO. Symbol * was designated each of the three codons ter of (1).

inevitable blocking that take place in all the encodings of the amino acids contained in these pairs. The second integral characteristic of the genetic code [20] containing two components was introduced into consideration:

$$q : q_{\min} = 31, q_{\max} = 210 \tag{6}$$

With this in mind, in addition to the unavoidable, the possible locks that arise for a limited number of encodings were introduced; these locks are the main component of 210 locks, which also include 31 inevitable locks (see Figure 8 and Table 3).

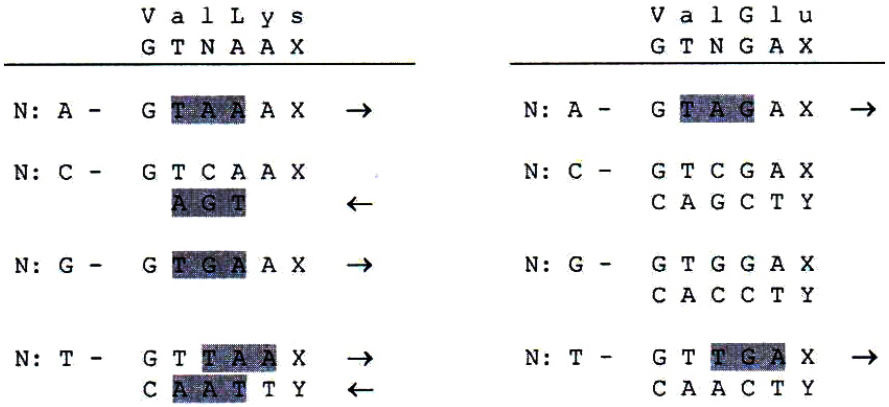

Figure 8. The inevitable lock for a pair of amino acids ValLys (left) and the who-possible blocking for a pair of amino acids ValGlu (right). The ter codons of (1) are shaded. Arrows indicate reading direction. The full list of inevitable locks is presented in Table 3.

Table 3. Complete list of amino acid pairs that cause an inevitable block (column 1), with specification of the numbers of RFs that are blocked (column 2).

№	1	2	№	1	2
1	MetMet	2	17	IleMet	2, 5
2	MetAsn	2	18	ValMet	2, 5
3	MetLys	2	19	LeuMet	2, 4, 5
4	MetIle	2	20	IleAsn	1, 2, 5
5	MetThr	2	21	ValAsn	1, 2, 5
6	PheTyr	3	22	LeuAsn	1, 2, 4, 5
7	TyrTyr	3	23	IleLys	1, 2, 5
8	HisTyr	3	24	ValLys	1, 2, 5
9	AsnTyr	3	25	LeuLys	1, 2, 4, 5
10	AspTyr	3	26	IleIle	2, 5
11	CysTyr	3	27	ValIle	2, 5
12	PheMet	5	28	LeuIle	2, 4, 5
13	PheAsn	1,5	29	IleThr	2, 5
14	PheLys	1,5	30	ValThr	2, 5
15	PheIle	5	31	LeuThr	2, 4, 5
16	PheThr	5			

Column 2 of this table lists the RF numbers for which (potential) a lock may occur. The number of such RF, depending on the pair, varies from 1 to 4. Let's imagine the set of these pairs as two subsets. The first will include the pair, inevitably blocking the same RF. There were only 16 such pairs—these are the first 16 pairs from the Table 3. The ima eat block inevitably RF2 5 pairs of amino acids, which coincide with the set of (2) defined above; since RF2 is formed in a ter codon is TGA. For RF3 we have 6 blocking pairs matching the set (3), and in RF3 we form one of the codons ter: TAA or TAG. Of particular note is the lock for 5 pairs of numbers 12 - 16, which coincide with pairs of (4) defined above. Only note that each of these latter pairs will inevitably inhibits RF5, as RF5 formed one of the kodon ter: TAA or TGA .However, in PheAsn, PheLys pairs in addition to RF5 in Table 3 is also indicated by RF1. However, the latter RF does not correspond to the inevitable blocking, unlike RS5. Thus, pairs 1 - 16 of Table 3 form a set of amino acid pairs forbidden to overlap the two genes and have been established above. Previously, a numerical characteristic was introduced, which was indicated by the letter p, and which corresponds to the number of different blocking pairs from (2)-(4), we have a value of p from (5.) Thus, the utilization of a subset of the inevitable blokirouac allows the connection of the studied characteristics: the inequality:

$$0 \le p \le q_{\min} \tag{7}$$

In other words, the integral characteristic of the genetic code p is not independent, but is determined by the choice of the characteristic $q_{\min}$, which is used in solving a completely different problem—in blocking non-overlapping genes.

When considering only one problem-overlapping pairs of genes, it could be concluded that the genetic code was "chosen" for overlapping genes, since only 16 pairs out of 400 possible pairs are suitable for overlapping. This is true for all 5 methods of pair re-discovery of genes, permitted by the structure of DNA. However, when considering two problems and two integral characteristics p and q, it was found that the genetic code was focused on the "choice" of the two-component integral characteristic (6), one of the components of which according to the inequality (7) determines the area of the value of the other integral characteristic p. Thus, the smallness of the integral characteristic of p is a consequence of a more general principle associated with the selection of the whole set of inevitable amino acid pairs that create blockers. I mean… amino acid pairs corresponding to the characteristic p can be "selected" only from this limited set of the corresponding $q_{\min}$, and not from the complete set of 400 amino acid pairs. According to what criterion the proposed "choice" of the genetic code has gone out yet remains non-existent.

In connection with the analysis of the blocking problem, we have investigated a number of genomes with a total number of genes more than 200,000. The story of this work requires a separate consideration; a more detailed analysis is presented in [21]. Note only one result obtained for the human genome. Of the 25,613 genes in this genome, three genes do not contain any blockages: for each

of them, the figures are similar to rice.8 do not contain any codon of termination in any of the five alternative RS. It MT1M and MT1G genes from chromosome 16 and KLK8 of chromosome 19 (see Figure 9).

In connection with the obtained result, a number of hypotheses were put forward, from which the simplest is nothing more than cases of overlap of 6 genes.

11. Mathematical Analysis of Code Deviance

The task of studying locks for deviant codes allowed to complete the analysis of one of the most important fundamental problems related to the role of all the reinterpreted codons. Data were obtained for mtDNA of two organisms: H. Sapiens (code K1) and A. Mellifera (code K2). Refer to the Table 4.

It follows from the table that for mtDNA H. Sapiens (K1) there is a participation of all the reinterpreted codons (ATA(Ile) → Met, TGA(ter) → Trp, AGA(Arg) → ter, AGG(Arg) → ter) in the blocking process, as well as the

Figure 9. KLK8 gene from human genome and 5 alternative RF. Each of these RF does not contain a single termination codon, compare with Figure 7. Here, the standard one-letter coding is used for amino acids.

Table 4. Summary table of the participation of reassignment codons in two functions: in the lock (in column 1, the sign + corresponds to participation, the sign-to-non-participation), or in the genetic overlap (column 2). Data were obtained for mtDNA of two organisms: H. Sapiens (code K1) and A. Mellifera (code K2). Rethinking codons are indicated in the column: deviations from the standard code.

Organism	Deviations from standard code	1	2
H. Sapiens (K^1)	ATA(Ile) → Met	+	−
	TGA(ter) → Trp	+	+
	AGA(Arg) → ter	+	−
	AGG(Arg) → ter	+	−
A. Mellifera (K^2)A	ATA(Ile) → Met	+	+
	TGA(ter) → Trp	−	+
	AGA(Arg) → Ser	+	−
	AGG(Arg) → Ser	+	−

re-interpreted TGA(ter) → Trp codon is also involved in gene overlap, forbidden for the standard code. For mtDNA A. Mellifera (K2), all the reinterpreted codons(AGA(Arg) → Ser, AGG(Arg) → Ser, ATA(Ile) → Met) participate in the blocking process, in addition, two such codons participate in the closures: TGA(ter) → Trp and ATA (Ile) → Met. Thus, we have shown that all the re-interpreted codons in each of these two deviant codes were used either in the process of blocking or in the process of overlapping genes (of course, prohibited for the standard code), or in both processes.

12. Is the Code Arbitrary?

The obtained results lead to the conclusion that the code deviations from the standard one are not of a random nature, but bear a very clear functional load (cf. "codon Reinterpretation indicates that random changes can occur in the genetic code of mitochondria", see [22]). In the last monograph we also read "the code seem to have been selected arbitrarily…" ("the Code, apparently, was 'you-bran' arbitrarily…"). From these results, it follows that it is possible for all semantic codon families to record two protein sequences almost without interference with the same DNA region, and for this, the most favorable (by the combination of amino acids in the overlap) one of the 5- and the variants of such a compact record of genes (5 cases of overlap) can be used. There is a categorical prohibition for no more than 5% of amino acid pairs, both for the standard code and for all 14 known non-standard codes. I mean… 15 code tables satisfy the same General property. This leaves no chance for any arbitrariness.

13. The Sets of Elementary Overlapping

The main working sets in this theory are the sets of elemental genetic overlaps, which are presented for the first time on pages 13 - 27 in [23], and the examples are given in **Figure 10**. Elementary overlapping is overlapping for single amino acids.

Such complete sets have been used repeatedly in the course of the construction of this theory. First, in the proof of the theorem for the genetic code and then these sets were modified 14 times (by the number of deviant codes) to obtain the first integral character of these 14 codes, which are presented in the **Table 2**. The most important stage of the research was connected with the mathematical analysis of ambiguities in these sets the Components of the study are numerous elementary genetic overlappings is overlapping for single amino acids. The analysis showed that the set contains features that were called ambiguities. The investigated ambiguities correspond to the cases when for the same pair of amino acids there is more than one elementary overlap. Like all special cases in mathematics, this phenomenon has attracted our attention. It is important to note that the results obtained are applicable to the whole diversity of wildlife, whose proteins are recorded by almost the same genetic code. The analysis showed that the ambiguities occur only in cases of overlapping genes belonging

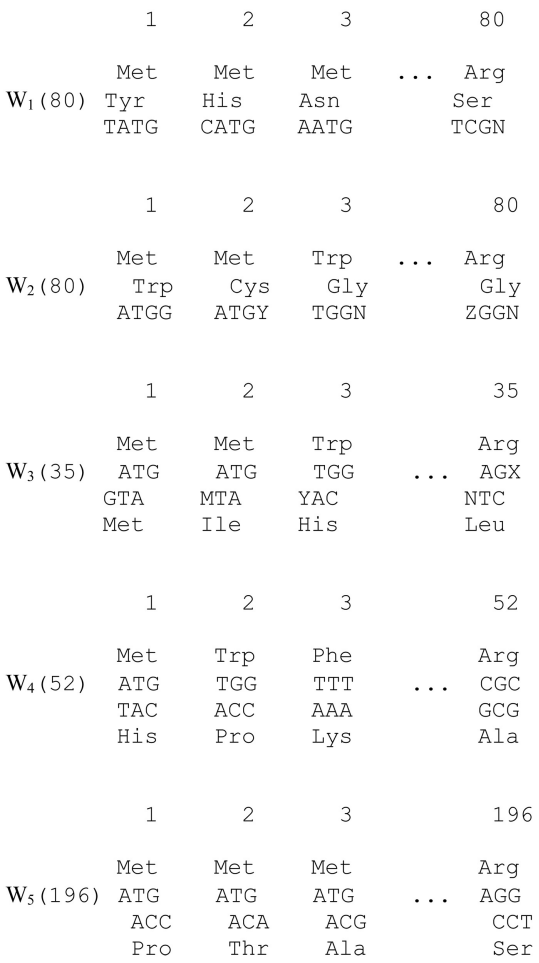

Figure 10. Some elementary overlaps from five sets corresponding to five cases of possible overlaps of gene pairs from **Figure 4**. Symbols N: A, C, T, G; M: A, T, C; X: A, G; Y: T, C; Z: A, C.

to different DNA chains. The complex numbers, the elementary beams was relatively small—only 6. The study revealed three functions of the possible use of these ambiguities. These functions were three [24]. One of the functions of ambiguity that is—has been succeeded in the new model proposed by the author of the wound. It consists in the fact that the overlapping pairs of genes belonging to different DNA chains are mathematical analogues of the stems of the secondary structure of matrix RNA. It is shown that due to the ambiguity it is possible to "regulate" the value of free energy of the stem functionally significant biochemical characteristics [25]. Now about other two functions, it is clear [24]. The first of them is related to the solution of the problem of potential positions of silent mutations for cases of gene overlap belonging to different DNA chains. The second is related to the expansion of the possibility of constructing sets of genetic overlaps of more than two genes; the structure of possible overlaps of 6 genes in the human genome is analyzed.

The study of the spatial structure of DNA showed that in addition to three

families of forms of double helices with antiparallel orientation of the threads, it is possible to form double helices of DNA with a parallel orientation of the threads. Mathematical analysis of such cases is given in section 4.4 of [26], where we analyzed all three new cases of the overlapping pairs of genes.

14. Conclusions

On the basis of the constructed theory, one of the breakthrough problems—the problem of calculating the genetic code—was solved. Such tasks in the world are unknown and could be set only in the 21st century. One of the approaches to solving this problem is given in article [27]. The mathematical theory of the genetic code is constantly evolving. We would like to point out the last two works in this direction.

Mathematical analysis of large genomes is an actual problem in connection with the development of genome decoding methods. By now, the genomes of man and some other organisms have already been decoded. The paper presents a numerical analysis of some characteristics of the genetic code common to all these genomes. The obtained results allow us to formulate a new property of the genetic code for the overlap of 6 and 3 genes from one DNA chain: the choice of three terms—toric codons from 64 possible triples of the genetic code has virtually no effect on the power of nucleotide chain sets, allowing six-fold or three-fold overlap of genes [28].

The second article [29] is connected with the mathematical analysis, which allowed formulating the property of the three terminator codons of the standard genetic code, when compared with other theoretically possible triples. The mathematical analysis allows formulating the following property of the three terminator codons of the standard genetic code (these codons stop white synthesis with DNA and do not encode any amino acid) when compared with other theoretically possible triples. For any choice of three terminator codons, one can find a DNA chain with a length of 11 nucleotides, where translation is completely forbidden. For the three term-end codons of the standard genetic code on any DNA chain of length 10 or less nucleotides, at least one reading frame is open, *i.e.* the translation process for at least one of the three reading frames is possible. This length is the maximum possible and when you select another three terminator codons, it may be less. For a triplet of terminator codons of a standard genetic code on any DNA strand of length 10 or less nucleotides, at least one reading frame is open, *i.e.* a translation process is possible for at least one of the three reading frames. This length is the maximum possible and may be less when choosing another triple of terminator codons. The number of terminator triples with such properties is 2280, and the probability of falling into this group of randomly selected triples of terminator codes is less than 0.06.

Acknowledgements

The author thanks a brilliant interpreter O. N. Kozlova, who translated this text from Russian.

Funding

The work was supported by Russian Foundation for Basic Research (project codes16-01-00018, 17-01-00053).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ycas, M. (1969) The Biological Code. North-Holland Publishing, Amsterdam, London, 359 p.

[2] Ychas, M. (1994) Meaning and Mechanisms.

[3] Mendel, G. (1866) Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, **4**, 3-47.

[4] Schrödinger, E. (1944) What Is Life? The Physical Aspect of the Living Cell. University Press, Cambridge.

[5] Watson, J.D. and Crick, F.H.C. (1953) A Structure for Deoxyribose Nucleic Accid. *Nature*, **171**, 737-738. https://doi.org/10.1038/171737a0

[6] Watson, J.D. (1968) The Double Helix. A Personal Account of the Discovery of the Structure of DNA. Atheneum, New York.

[7] Lewin, B. (1997) Genes VI. Oxford University Press, Oxford, 879 p.

[8] Barrell, B.G., Air, G.M. and Hutchisoп III, C.A. (1976) Overlapping Genes in Bacteriophage ΦX174. *Nature*, **264**, 34-41. https://doi.org/10.1038/264034a0

[9] Kozlov, N.N. (1999) Involvement of Each of 64 Codons in Gene Overlappings. *Doklady Biochemistry*, **367**, 126-128.

[10] Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M. and Smith, M. (1978) The Nucleotide Sequence of Bacteriophage ΦX174. *Journal of Molecular Biology*, **125**, 225-246. https://doi.org/10.1016/0022-2836(78)90346-7

[11] Nakayama, T., Asai, S., Takahashi, Y. and Nishida, Y. (2007) Overlapping of Genes in the Human Genome. *Nevill Juvenile Bonfire Society*, **3**, 14-19.

[12] Guyader, M., Emerman, M., Sonigo, P., Clavel, F., Montagnier, L. and Alizon, M. (1987) Genome Organization and Transactivation of the Human Immunodeficiency Virus Type 2. *Nature*, **326**, 662-669. https://doi.org/10.1038/326662a0

[13] Barrell, B.G., Bankier, A.T. and Drouin, J. (1979) A Different Genetic Code in Human Mitochondria. *Nature*, **282**, 189-194. https://doi.org/10.1038/282189a0

[14] Kozlov, N.N. (2014) One Integral Characteristic of the Set of Genetic Codes. The Property of All Known Natural Codes. *Mathematical Models and Computer Simulations*, **6**, 622-630. https://doi.org/10.1134/S2070048214060064

[15] Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. and Ysebaert, M. (1976) Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene. *Nature*, **260**, 500-507. https://doi.org/10.1038/260500a0

[16] Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin,

J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R. and Young, I.G. (1981) Sequence and Organization of the Human Mitochondrial Genome. *Nature*, **290**, 457-464. https://doi.org/10.1038/290457a0

[17] Clary, D.O. and Wolstenholme, D.R. (1985) The Mitochondrial DNA Molecule of Drosophila Yakuba: Nucleotide Sequence, Gene Organization, and Genetic Code. *Journal of Molecular Evolution*, **22**, 252-271. https://doi.org/10.1007/BF02099755

[18] Cantatore, P., Roberti, M., Rainaldi, G., Gadaleta, M.N. and Saccone, C. (1989) The Complete Nucleotide Sequence, Gene Organization, and Genetic Code of the Mitochondrial Genome of *Paracentrotus lividus*. *The Journal of Biological Chemistry*, **264**, 10965-10975.

[19] Crozier, R.H. and Crozier, Y.C. (1993) The Mitochondrial Genome of the Honeybee *Apis mellifera*: Complete Sequence and Genome Organization. *Genetics*, **133**, 97-117.

[20] Kozlov, N.N. (2011) Integral Characteristics of Genetic Code. *Mathematical Models and Computer Simulations*, **3**, 123-134. https://doi.org/10.1134/S2070048211020050

[21] Kozlov, N.N. (2013) Some New Characteristics of Large Genomes. *Mathematical Models and Computer Simulations*, **5**, 220-228. https://doi.org/10.1134/S2070048213030071

[22] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. (1994) Molecular Biology of the Cell. Gorland Publishing, Inc., New York, London, 1294 p.

[23] http://www.keldysh.ru/papers/2004/prep64/prep2004_64.html

[24] Kozlov, N.N. (2015) Three Function Ambiguity from the Sets Generated by the Genetic Code. *Mathematical Models and Computer Simulations*, **7**, 401-408. https://doi.org/10.1134/S2070048215050063

[25] Kozlov, N.N. (2013) One Function of Ambiguities from the Sets Generated by the Genetic Code. *Mathematical Models and Computer Simulations*, **5**, 17-24. https://doi.org/10.1134/S2070048213010067

[26] Kozlov, N.N. (2014) Genetic Code: A Mathematician's Point of View. Palamarium Academic, Hamburg, 336 p.

[27] Kozlov, N.N. (2017) Computation of the Genetic Code: Full Version. *Journal Computer and Communications*, **5**, 78-94. https://doi.org/10.4236/jcc.2017.510008

[28] Kozlov, N.N., Kugushev, E.I. and Eneev, T.M. (2017) Genetic Code Potential for Overlaps of Six and Three Genes. *Doklady Mathematics*, **95**, 161-163. https://doi.org/10.1134/S1064562417020168

[29] Kozlov, N.N., Kugushev, E.I. and Eneev, T.M. (2017) Mathematical Analysis of Codons That Stop Protein Synthesis. *Doklady Mathematics*, **96**, 571-573. https://doi.org/10.1134/S1064562417060102