ISSN 2070-0482, Mathematical Models and Computer Simulations, 2014, Vol. 6, No. 6, pp. 622–630. © Pleiades Publishing, Ltd., 2014.
Original Russian Text © N.N. Kozlov, 2014, published in Matematicheskoe Modelirovanie, 2014, Vol. 26, No. 5, pp. 113–125.

One Integral Characteristic of the Set of Genetic Codes. The Property of All Known Natural Codes

N. N. Kozlov

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

Received January 30, 2013

Abstract—Earlier, this author introduced the integral characteristics of the genetic code ("Integral characteristics of the genetic code," Mathematical Modeling, vol. 22, no. 9, 2010). One of these characteristics that is correlated to the potential of a code for building overlapping genes, when the same piece of DNA encodes two protein sequences, is considered here. This is an investigation of a variety of genetic codes that corresponds to two groups of such codes. First of all, the hypothetical codes were considered and this has allowed the establishment of a range of changes in this characteristic by the use of different numbers of codon permutations in the standard genetic code. The second group of codes is the natural genetic codes. It has turned out that all of the known natural codes today (currently, 15) have one common property. This property is formulated. Note that the first natural standard code was found in the human cell in 1979, in a separate organelle—in mitochondria.

DOI: 10.1134/S2070048214060064

Only pairs of genes interconnected with each other are analyzed here. Such genes are called overlapping and they were first discovered in 1976. For the first time the case of coding two nonhomological proteins was found, whose genes were written with a shift for one nucleotide. It was the overlapping of two genes belonging to the same DNA chain. By now already all possible for DNA cases of overlapping gene pairs have been found; such cases are now not more than 5 (Fig. 1), in two of them (the upper strip in Fig. 1) only one DNA chain participates, in the three others (the bottom strip in Fig. 1), the overlapping genes correspond to different DNA chains.

The record of amino acid sequences is made based on the genetic code by the gene text and is given over such a text for the DNA plus-chain and below such a text for the minus-chain. Besides, because of the anti-parallelism of the DNA chains, the reading of these sequences is from left to right for the plus-chain and from right to left for the minus-chain (see the arrow in Fig. 1). The shifts shown in Fig. 1 denote the number of nucleotides onto which the corresponding genes are pushed: for overlaps in one DNA chain gene B_{12} is shifted for -1 nucleotide with respect to gene B_{11}, and gene B_{22} is shifted for $+1$ nucleotide to gene B_{21}. Figure 1 also shows 3 cases of overlaps of the pairs of genes from different DNA chains: shift -1 (B_{32} relative to B_{31}), shift 0 (B_{42} relative to B_{41}), and shift $+1$ (B_{52} relative to B_{51}).

The studies [1–4] in mathematical analysis of overlapping genes are related only to cases 1 and 2 shown in Fig. 3, i.e., to overlaps in one DNA chain. Based on the study of the experimental data in such overlaps, the interaction of a number of features of the observed overlaps with some features of the structure of the genetic code was found. In order to receive the basic results in such an interaction, the mathematical modeling of the entire set of these overlaps was investigated. It is shown [1], that for the most extended genetic overlap found in the GSHV virus, containing more than a half of the genome size, i.e., 428 codons [5], the potentially maximum possible number of overlaps is $\sim 10^{746}$. During the study, all possible sets of amino acids were found for which overlapping takes place. Overlapping includes such cases when in the gene corresponding to the shifted state not a single of the three ter codons can take place: TAA, TAG, TGA. None of these codons can be present in any amino acid sequence, because such a codon stops the protein synthesis. We call an overlap as total if it takes place for a sequence of any length with an arbitrary combination of amino acids. The mathematical modeling of complete sets of overlaps for cases 3–5 (see Fig. 1) that was performed by analogy with cases 1–2 has allowed the establishment of an earlier unknown property of the standard genetic code K^0 that is formulated here as a theorem [5].

Theorem. *The standard genetic code tolerates the total overlapping for each case out of* 1–5 *cases except for sequences containing at least one of the pairs of amino acids in the following three overlapping cases*:

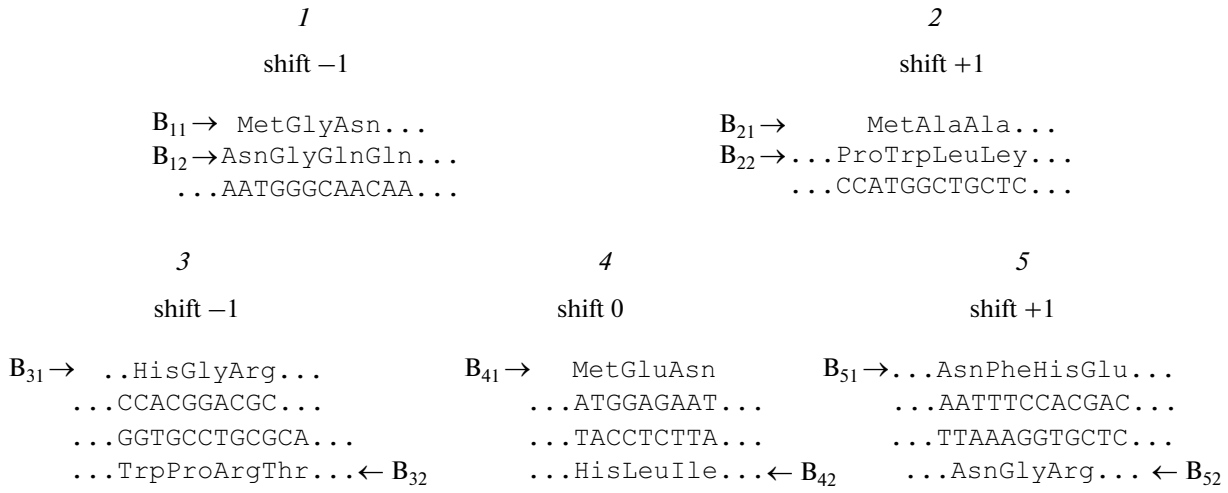

Fig. 1. Five possible cases of overlapping genes corresponding to one (*1*, *2*), or two DNA chains (*3*–*5*). These are illustrative fragments of gene overlaps which in reality can reach sizes of over 1000 nucleotides. Reading of texts in this is in different direction (shown by the arrow): from left to right for B_{11}, B_{12}, B_{21}, B_{22}, B_{31}, B_{41}, B_{51} and from right to left for B_{32}, B_{42}, B_{52}.

in the case of 2 it is 5 pairs:

MetMet, MetAsn, MetLys, MetIle, MetThr, (1)

in the case of 3 it is 6 pairs:

PheTyr, TyrTyr, HisTyr, AsnTyr, AspTyr, CysTyr, (2)

in the case of 5 it is 5 pairs:

PheMet, PheAsn, PheLys, PheIle, PheThr. (3)

Based on this theorem, we introduce into the consideration the numerical characteristic, call it integral, and denote it through p. This is one of two integral characteristics of the genetic code introduced by us earlier [6]. Such a characteristic corresponds to multiple bonds in genetic overlaps and points to all the sets of pairs of amino acids forbidden for overlaps of genes in all of the five mentioned cases of overlapping. Thereby, according to the theorem, the integral characteristic for the standard code, p^0, is 16. For codes deviating from the standard code, this characteristic p can change. Below we consider such changes. The study is for the set of genetic codes corresponding to two groups of such codes.

First of all, the hypothetical codes were considered, which made it possible to establish a range of changes of this characteristic by use of a different number of codon permutations in the standard genetic code. For this study, the sets of elementary genetic overlaps (e.os) were employed. Such e.os correspond to overlaps for solitary amino acids and were calculated separately for each out of five cases of overlaps. As a result five such sets, W_1–W_5, were found (see the release version in [7]) for the first time in [8], where the index denotes the case of overlapping. The total number of e.os in all the sets proved to be 443. Besides, use is made of the notion of the reading frame (RF). There are five such frames: RF1–RF5, and they correspond to sequences arising in shifts corresponding to 1–5 cases of overlaps.

The second group of the below codes analyzed is natural genetic codes. The number of the latter codes is 15. Note that the first natural substandard code was found in 1979 in a human cell, in a separate organelle in mitochondrion.

INTEGRAL CHARACTERISTIC FOR HYPOTHETICAL CODES

We consider three sets of hypothetical codes constructed based on the mathematical analysis of e.o sets in order to change the characteristic p compared to the value $p^0 = 16$. The theoretically possible range of change of this integer characteristic p ranges from 0 to 400. The choice of the used changes is based on the structure of the e.o sets.

Consider at first the set of K_{11}–K_{16} codes from Table 1, each able to be formed using not more than one permutation. Codes K_{11}–K_{13} correspond to permutations among the sense codons: TGC(Cys) → Trp for K_{11}, CAC(His) → Gln for K_{12}, and ATA(Ile) → Met for K_{13}. Codes K_{14}–K_{16} correspond to permutations of the sense codon into the set ter, i.e.; the set ter is expanded to four: ATA(Ile) → ter for K_{14} (changes for

this code are given in the square brackets in the third column), GGC(Gly) → ter for K_{15}, CGA(Arg) → ter for K_{16}. Each of the codes K_{11}–K_{16} corresponds to the meaning of the characteristic $p > p^0$. It is such permutations that have been chosen (the total number of such single permutations for K^0 is over 100; this number is based on the analysis of the structure of the sets W_1–W_5).

In code K_{11}, additionally to 16 pairs for K^0, another six blocking pairs of amino acids arise. This takes place for four pairs of amino acids (in Table 1 the codons from the ter set are shown in bold) CysAsn, CysAsp, CysLys, CysGlu (PC1 blocking takes place, see the first column in Table 1, where the PC number is shown on the right), as well as for the two pairs CysHis and CysGln (PC3 blocking proceeds). The blocking is created by two ter codons—they are TAA and TGA. Thus, for K11 we have $p = 16 + 6 = 22$. This value $p = 22$ corresponds also to codes K_{12}–K_{14}. For K_{12} (CAC(His) → Gln) PC1 blockings also arise because of ter: TGA. This takes place for four pairs of amino acids: HisAsn, HisAsp, HisLys, HisGlu, and PC 3 blockings for two pairs of amino acids: HisGln and HisHis.

The case of two codes K_{13}, K_{14} stands by itself. For K_{13} (ATA(Ile) → Met) and K_{14}(ATA(Ile) → ter) the same additional blockings arise due to ter codons, belonging to K^0. Due to TAX, X: A, and G, PC 2 for the pair IleTyr is blocked, and due to TXA, PC 5 is blocked for the pairs IleMet, IleAsn, IleLys, IleIle, and IleThr. This means that for such codes additional blockings arise only for PC2 and PC5. It should be also noted that the blocking pairs of amino acids for K_{13}, connected with the expansion of the Met codings up to two codons persist according to (1). In cases K_{15} and K_{16} additional blockings appear due to the rethought codons GGC y K_{15} and CGA y K_{16}, which, accordingly, are included into the set ter. For $_{15}$(GGC(Gly) → ter) there are additional blockings: PC1 for pairs MetAla and TrpAla; PC2 for pairs TrpHis, TrpGln, and TrpPro; and, finally, blockings PC 3 for pairs MetPro and TrpPro. All these blockings are created by the codon ter: GGC. By means of one permutation there is a maximum increase of p for the case K_{16} (CGA(Arg) → ter): from $p^0 = 16$ to $p = 28$. Then ter: YGA, Y: T,C additionally blocks only PC 1 for 12 pairs: PheAsp, TyrAsp, HisAsp, AsnAsp, AspAsp, CysAsp, PheGlu, TyrGlu, HisGlu, AsnGlu, AspGlu, and CysGlu.

Turn now to hypothetical codes K_{21}–K_{26} (Table 2), which were formed by such permutations that for each of them we have the absolute minimum of the possible value $p - p = 0$. Parameter ν is the number of codon families of the code which diverts the structure from the regular one (in each three-letter coding for such a structure the first and second positions are the same) are given in brackets in the first line of the table. For K^0 we have $\nu^0 = 4$ (this is because of the irregularities for the families of codons Leu, Ser, Arg, ter), for K_{21}–K_{26} this number changes within 0 to 4.

We consider K_{21}. This is the only code with characteristic $p = 0$, which can be formed of K^0 by permutation of only one codon. This statement was first formulated in [5]. The codes K_{22}–K_{26} are given to present changes associated with the irregularity of the codes. The accepted regularity for Ser, Leu, and Arg (y K_{22} and K_{23}) or for Leu (y K_{24}) has been achieved due to the irregularities for Tyr (y K_{22}) or for Met and Tyr (y K_{23} and K_{24}). Besides, for K_{22}–K_{24} the set ter coincides with this value for K^0. For codes K_{25} and K_{26}, the same one and only one value of the set ter: TAG is accepted, and for K_{25} we have $\nu = 3$, and for K_{26} there is $\nu = 0$. Or the regularity of the code K_{26} is achieved due to reduction of the set ter down to the only value TAG.

Finally, the problem of the search for hypothetical codes was stated, in which the characteristic p is far greater than the value of p^0. In doing this, the solution was being sought for not among all possible sets of codon families but with account for the structures of natural substandard codes, see in detail below. For such codes the set of ter codons has no more than four codons, the set of sense codons for one amino acid is not more than eight codons. Three out of the series of the obtained solutions are given in Table 3.

Now turn to Table 3. $p = 58$ corresponds to code K_{31}, and the difference from K^0 takes place in three codon families: Trp, Cys, and ter, and we have $\nu = 3$. The reduction of this value compared to $\nu^0 = 4$ is because of the regularity of the set ter: TGN, N: A, C, T, and G. For K_{32} with the same set ter, we have $p = 155$ or almost an order higher $p^0 = 16$, and $\nu = \nu^0 = 4$. This increase in p is achieved due to the change of 12 codon families in K^0. For code K_{33} the accepted set ter corresponds to K^0, and the value $p = 83$, $\nu = 6 > \nu^0$ was obtained.

Tables 1–3 illustrate only some of the more than 100 hypothetic codes which were investigated. The range of the change of the characteristic p: from 0 (this is the absolute min p) up to 155 (about a 10-fold increase compared to p^0). The mathematical analysis of hypothetic codes is very important in connection with the continuous publications of new natural substandard genetic codes. Besides, recently the experimental works have been started on the creation of some versions of the genetic code that deviates from K^0.

Table 1. Additional blockings on genetic overlaps for cases of 1–5 overlaps (see the figure on the right of the overlaps) in hypothetical codes K_{11}–K_{16}. Each of these codes is formed by a permutation in K^0 of only one sense codon

K_{11} ($p = 22$)	K_{12} ($p = 22$)	K_{13} [K_{14}] ($p = 22$)	K_{15} ($p = 23$)	K_{16} ($p = 28$)
Permutations of only one codon				
GC(Cys) → Trp Cys:TGT	CAC(His) → Gln His:CAT	ATA(Ile) → Met Met : ATX, Ile:ATY [ATA(Ile) → ter]	GGC(Gly) → ter	CGA(Arg) → ter
CysAsn 1 TG**TAA**Y	HisAsn 1 CA**TAA**Y	IleTyr 3 ATYTAY TA**X**A**T**X	MetAla 1 AT**GGC**N	PheAsp 1 TT**YGA**X
CysAsp 1 TG**TGA**Y	HisAsp 1 CA**TGA**Y	IleMet 5 ATYATX	TrpAla 1 TG**GGC**N	TyrAsp 1 TA**YGA**X
CysLys 1 TG**TAA**X	HisLys 1 CA**TAA**X	T**X**T**A**Y	TrpHis 2 T**GGC**AY	HisAsp 1 CA**YGA**X
CysAsn 1 TG**TGA**X	HisGlu 1 CA**TGA**X	IleAsn 5 ATYAAY T**X**T**T**X	TrpGln 2 T**GGC**AX	AsnAsp 1 AA**YGA**X
CysHis 3 TGTCAY AC**AGT**	HisHis 3 CATCAT GT**AGT**A	IleLys 5 ATYAAX T**X**T**T**Y	TrpPro 2 T**GGC**CN	AspAsp 1 GA**YGA**X
CysGln 3 TGTCAX AC**AGT**	HisGln 3 CATCA GT**AGT**	IleIle 5 ATYATY T**X**T**A**X	MetPro 3 A TGCCN TA**CGG**	CysAsp 1 TG**YGA**X PheGlu 1 TT**YGA**X
		IleThr 5 ATYACN T**X**T**G**	TrpPro 3 TGGCCN AC**CGG**	TyrGlu 1 TA**YGA**X
				HisGlu 1 CA**YGA**X
				AsnGlu 1 AA**YGA**X
				AspGlu 1 GA**YGA**X
				CysGlu 1 TG**YGA**X

The termination codons are shown in bold. For K_{13} and K_{14}, the solutions coincide.

Table 2. Standard K^0 and hypothetical codes K_{21}–K_{26}, for each of which the characteristic $p = 0$. Only data of the family deviating from the genetic code from its standard structure are shown. Denotations N: T, C, A, G; M: T, C, A; X: A, G; Y: T, C

		K^0	(4)	K_{21}	(4)	K_{22}	(2)	K_{23}	(3)	K_{24}	(4)	K_{25}	(3)	K_{26}	(0)
1	Met	(1)	ATG			(4)	ATN	(3)	ATG, AGY	(2)	ATG, ATT			(4)	ATN
2	Trp	(1)	TGG									(2)	TGX	(2)	TGX
3	Phe	(2)	TTY			(4)	TTN	(3)	TTN	(3)	TTM			(4)	TTN
4	Tyr	(2)	TAY	(3)	TAY, TGA	(4)	TAY, AGX	(3)	TAY, AGX	(3)	TAY, TTG	(3)	TAM	(3)	TAM
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			(2)	AGY			(2)	ATZ			(4)	AGN
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			(4)	TCN	(4)	TCN					(4)	TCN
19	Leu	(6)	CTN, TTX			(4)	CTN	(4)	CTN	(4)	CTN			(4)	CTN
20	Arg	(6)	CGH, AGX			(4)	CGN	(4)	CGH					(4)	CGN
	ter	(3)	TAX, TGA	(2)	TAX							(1)	TAG	(1)	TAG

In particular, expansion of the number of the amino acids coded in E. coli was obtained and it is noted [9] that the developed approach can be laid on the basis of the method for expansion of the genetic repertoire of live cells and building-in the amino acids with new structural, chemical, and physical properties into proteins.

THE PROPERTY OF ALL KNOWN NATURAL CODES

In addition to the role established above of the rethought codons in genetic overlaps, it has turned out that for all the natural genetic codes (known to date) there is one common property. According to the data on the Internet (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t), the total number of such codes is at present not more than 20. In this study, only those codes were identified for which every codon is meaningful; i.e., the codes where participation of some codons in codings was not found were excluded. Also, the deviant codes with a structure coinciding with the structure K^0, but with different codons of initiation were excluded. As a result, no more than 14 substandard codes remained. The p characteristics for each of the mentioned codes are shown in Table 4.

The calculations have shown that for no more than one of the codes, K^{14} (Thraustochytrium Mitochondrial Code), we have $p^{14} = 21$; for all the remaining 13 deviant codes, the value p does not exceed the value $p^0 = 16$. It has also turned out that for one of the codes, K^9 (The Alternative Flatworm Mitochondrial Code), we have $p^9 = 0$. This value corresponds to the absolute minimum equal to zero for the characteristic p.

Table 3. The standard code K^0 and hypothetical codes K_{31}–K_{33}. These codes were sought in order to increase the integral characteristic p from 16 to 58, 83 and, finally, to 155 (K_{32})

		K^0	K_{31}	K_{32}	K_{33}
1	Met	(1) ATG			(2) ATX
2	Trp	(1) TGG	(1) TAG	(1) TTG	(1) GGT
3	Phe	(2) TTY		(1) CTG	(4) GAN
4	Tyr	(2) TAY		(1) GTG	
5	His	(2) CAY		(1) CAT	(1) GTA
6	Asn	(2) AAY			
7	Asp	(2) GAY			(8) GZN_T, GTZ
8	Cys	(2) TGY	(1) TAA	(1) CAC	(4) AGN
9	Gln	(2) CAX		(1) CAG	(6) CTY, CGY, CCY
10	Lys	(2) AAX			
11	Glu	(2) GAX		(1) CAA	(2) GYT
12	Ile	(3) ATM			(2) TYA
13	Val	(4) GTN		(5) GTM, GAX	(2) ATY
14	Pro	(4) CCN		(8) CCN, TAN	(1) CTA
15	Thr	(4) ACN			(1) ACT
16	Ala	(4) GCN			(1) ACC
17	Gly	(4) GGN			(4) CAN
18	Ser	(6) TCN, AGY		(4) AGN	(2) ACX
19	Leu	(6) CTN, TTX		(6) CTM, TTM	(6) TTY, TGY, TCY
20	Arg	(6) CGN, AGX		(8) CGN, TCN	(8) CTG, ZZX, TN_AG
	ter	(3) TAX, TGA	(4) TGN	(4) TGN	
	p	16	58	155	83

Earlier it was shown [10, 11], that this minimum arises for the hypothetical code after a single TGA → Tyr permutation (see code K_{21} from Table 2). In the case K^9, we have five rethought codons, and it is possible to show that the possible number of codes for which $p = 0$ with such a number of reconsiderations is rather high. In [5] a set of codes with the zero characteristic p at no more than two reconsiderations is described.

Our study has shown the success of the introduction of the integral characteristic of the genetic code p. For all 15 known natural codes the value p does not exceed 21, or no more than about 5% of pairs (21 out of 400 possible) are disallowed. The average values of the characteristic p for the natural codes of Table 4 will be about 9, or almost half that the number for K^0.

Thus, an important property for all the 15 known natural codes was established. It is that each such code corresponds to the domain of small values of the characteristic p. From this it follows that the point-blank prohibition takes place for not more than about 5% of possible pairs of amino acids for all the five ways of paired genetic overlaps. The performed analysis affords ground to assume that in the further genetic experiments natural codes corresponding to this property can be found.The established property has a meaning for genetic codes created artificially. These works have been started relatively recently and have already given the first results [9]. It is of interest to find out how the behavior of the code with characteristic $p \geqslant 21$ will work. As to the hypothetical codes, such codes were investigated by us for the range of p from 0 to 155 (see above).

Table 5 provides some additional data for natural codes from Table 4 that expand our notions about other properties of substandard codes.

In Table 5, in addition to three sets (1)–(3), corresponding to K^0 for the deviant genetic codes K^1–K^{14}, no more than five changed sets of modifications of sets from (2) and (3) given below:

(2.1) is set (4.2) with the addition of the pair IleTyr;(2.2) is set (4.2) with the exclusion of the AsnTyr pair;(2.3) is set (4.2) with the addition of the CysTyr pair;

(2.4) is set (4.2) with the exclusion of the pair AsnTyr and addition of the pair AsnTyr;

Table 4. Standard K^0 and substandard 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(lle) → Met, AGX(Arg) → ter	7
K^2	The Invertebrate Mitochondrial Code	TGA(ter) → Trp, ATA(lle) → 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 Myco-plasma/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(lle) → Met, AGX(Arg) → G1y	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(lle) → Met, AGX (Arg) → Ser	6
K^{13}	Scenedesmus Obliquus Mitochondrial Code	TAG(ter) → Leu, TCA(Ser) → ter	10
K^{14}	Thraustochytrium Mitochondrial Code	TTA(Leu) → ter	21

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

Table 5. Pairs of amino acids unavoidably blocking PC1–PC5 for codes K^0–K^{14}

	Ter	PC1	PC2	PC3	PC4	PC5
K^0	ter^0	–	(1)	(2)	-	(3)
K^1	TAX, AGX	–	–	(2.1)	–	–
K^2	TAX	–	–	(2.2)	–	–
K^3	TAX	–	–	(2.2)	–	–
K^4	TAX	–	–	(2)	–	–
K^5	TGA	–	(1)	–	–	–
K^6	TAX	–	–	(2.3)	–	–
K^7	ter^0	–	(1)	(2)	–	(3)
K^8	TAX	–	–	(2.1)	–	–
K^9	TAG	–	–	–	–	–
K^{10}	TXA	–	(1)	–	–	(3)
K^{11}	TXA	–	(1)	–	–	(3)
K^{12}	TAX	–	–	(2.4)	–	–
K^{13}	TCA,TXA	–	(1)	–	–	(3)
K^{14}	TTA, ter^0	–	(1)	(2)	–	(3.1)

(3.1) is set (4.3) with the addition of five pairs: IleMet, IleAsn, IleLys, IleIle, and IleThr.

In PC3 blocking, in addition to the earlier introduced set (2) (in three codes: K^4, K^7, K^{14}), use is also made of the modified sets (2.1) (in two codes K^1, K^8) and sets (2.2) (in two codes K^2, K^3). For K^6 also set

K^{14} (Thraustochytrium Aureum)

IleLysPheIlePheTyr
LysIleTyrPhe *
ATAAAATTTATTTT**TTA**C
• •
17550

Fig. 2. The overlap in the DNA mitochondrion Thraustochytrium Aureum (substandard code K^{14}), in which the rethought codon TTA (*) takes part. The number below denotes the number of the T nucleotide in this genome in the overlap location.

(2.3) arises, and for code K^{12} set (2.4) arises. Each of these four sets has practically solid changes of pairs from (2). For PC5 blocking, in addition to set (3) (for four codes: K^7, K^{10}, K^{11}, K^{13}), for code K^{14} a blocking set (3.1) is used; it is set (3), increased by five above the mentioned additional pairs. Also note that for all the codes, except for K^8, the set of ter codons undergo changes. The most frequent rethinking in this is found for the TGA codon, which in eight codes becomes semantic. The number of ter codons changes from 1 (for K^5 and K^9) to 4 (for K^1 and K^{14}). Two codes (K^7 and K^{13}) have 3 ter codes each, and 8 codes 2 such codons each.

Table 5 presents sets of codons of the termination ter and blocking sets of pairs of amino acids. It is seen that none of the codes have pairs of amino acids blocking either PC1 or PC4. In other words, the natural reconsiderations are made in a very narrow domain, despite the fact that the number of the permuted codons may reach 5, as it takes place for the deviant codes K^9 and K^{12}. Our analysis of hypothetical codes has shown that even some single permutations may create additional blockings of PC1 or PC4. Thus, (see above) the permutation CGA(Arg) $\rightarrow$ ter increases the number of blocking pairs of amino acids to 28, and the additional (compared to K^0) 12 pairs correspond only to PC1. Another permutation CCA(Pro) $\rightarrow$ ter leads to the additional blocking of PC4. As is seen from Table 5, this kind of permutations were not used. So the blockings in K^0–K^{14} arise only in PC2 (this takes place for six deviant ones: K^5, K^7 , K^{10}, K^{11}, K^{13}, and K^{14}, and everywhere the same nonmodified set of pairs (1) is used), as well as in PC3 and PC5, where in addition to sets (2) and (3), their modifications were used.

The found property, as well as the insignificant share of genetic overlaps compared to the considerable majority of records of genes without overlaps, has led us to still another riddle. Why has Nature chosen the genetic codes with small values of the characteristic p? Further investigation has shown that the fact of the smallness of the bans on overlapping is also used for all the genes which are not overlapped. A publication on this problem is being prepared for publishing.

In conclusion, note that the code K^{14} is an exception to all the natural codes from Table 4; it is the only one of all the deviant codes that increases the introduced integral characteristic compared to the standard code due to the TTA codon, which results in the number of terminator codons increasing up to four. It would be also of interest to see this codon in operation and, especially, in the genetic overlaps. Figure 2 shows the genetic overlap written by this code, and this overlap does not seem possible for the standard code.

Note that the rethought codon TTA(Leu), which becomes the ter codon, corresponds to only one deviation of this deviant code K^{14} from the standard code. In Fig. 2 it is seen that this codon is used in order to overlap two genes from the single DNA chain and such a codon cannot be replaced by any of terminator codons of the standard code, which are also included in the structure of this deviant code. The economy of the size of this DNA is not more than five codons, and it seems impossible that due to this economy there could be a deviation of the code from the standard code. Facts of this kind are not sporadic and as a result have led to the search for not only this but also for another function which can be used by the rethought codons. The results of the performed study are now in print.

ACKNOWLEDGMENTS

The work was financially supported by the Russian Fund of Basic Research (project codes are 13-01-00133,11-01-00110).

REFERENCES

1. N. N. Kozlov, "To arbitrariness in the genetic code "choice", Dokl. Akad. Nauk **369** (4), 553–556 (1999).
2. N. N. Kozlov, "Analysis of the total set of overlapping genes," Dokl. Akad. Nauk **373** (1), 108–111 (2000).
3. N. N. Kozlov, "Mathematical analysis of overlapping genes and the structure of the genetic code," Mat. Model. **12** (7), 97–101 (2000).
4. N. N. Kozlov, "One way to store genetic information," Mat. Model. **14** (8), 72–78 (2002).
5. N. N. Kozlov, "Theorem for the genetic code," Dokl. Akad. Nauk **382** (5), 593–597 (2002).
6. N. N. Kozlov, "Integral characteristics of genetic code", Math. Models Comput. Simul. **3** (2), 123–135 (2011).
7. N. N. Kozlov, *Mathematical Analysis of Genetic Code* Kozlov (BIONOM, Moscow, 2010) [in Russian].
8. N. N. Kozlov, Preprint No. 64 (Keldysh IPM, 2004) (http://www.keldysh.ru/papers/2004/prep64/prep2004_64.html (MS Word).
9. Lei Wang, Ansgar Brock, Brad Herberich, and Peter G. Schultz, "Expanding the genetic code of Escherichia coli," Science **292** (5516), 498–500 (2001).
10. N. N. Kozlov, "Mathematical analysis of genetic codes," Mat. Biol. Bioinform. **1** (1), 70–96 (2006). http://www.matbio.org/downloads_en/Kozlov2006(1_70).pdf
11. N. N. Kozlov, "Overlapping genes and the structure of the genetic code," Preprint No. 113 (Keldysh IPM, 2005).

Translated by D. Shtirmer