

# The genetic code is not arbitrary

Nicolay Nicolaevich Kozlov

Keldysh Institute of Applied Mathematics. Russian Academy of Science, Moscow, Russia

**Abstract.** The study was carried out on the basis of mathematical analysis of experimental data: sets of genes and sets of natural genetic codes. All 5 cases of gene overlaps allowed by the DNA structure were investigated. Integral characteristics of genetic codes are introduced into consideration, which testify to the ability to overlap pairs of genes. It was possible to establish the functions in which all reinterpreted codons in the mitochondrial genetic codes (of humans and other organisms) participate, as well as a significant difference between the integral characteristics of such codes from the standard code. Our results allow us to conclude that code deviations from the standard carry a quite clear functional load. It follows from the above results that for all semantic families of codons, two protein sequences can be written that are practically not obstructed by the same DNA region, and for this you can use the most favorable (according to the combination of amino acids in the overlap) one of the 5 variants of such a compact notation of genes (5 cases overlap). A categorical ban exists for no more than 5% of amino acid pairs, both for the standard code and for all 14 known non-standard codes. Those. All 15 code tables have the same common property. This leaves no chance for any arbitrariness, chance. choose the structure of the genetic code. In the course of these studies, a mathematical theory of the genetic code has been developed, the first stage of research has already been published (see Appendix), a book of the results of the second stage has been prepared for publication, and work has been carried out on the third stage. It seems that this research topic of fundamental research is inexhaustible.

**Key words.** Genetic code, overlapping genes, deviant code, code possibilities for overlaps, common property of all known codes.

## 1. Foreword

After the successful completion of two labor-intensive and multi-year projects [1,2], I was allowed to conduct research in the field of modern genetics. Moreover, it was allowed to choose any task with one condition: it must be a discrete system with a large number of interconnected elements. This is how my leader, Academician T.M Eneev, set the task., who was one of the leading theorists of Russian cosmonautics and who would have been 100 years old in a year. The first line of new research was related to the problems of structuring RNA molecules, starting with the shortest ones, tRNAs [3]. Later, I turned to the analysis of special ways of recording genetic information - overlapping genes, when two proteins code for the same piece of DNA This effect was first experimentally established in 1976 when reading the first complete genome of bacterial viruses  $\Phi$ X 174 [4]. After these studies, their leader F. Sanger [5] became the only two-time Nobel Prize winner in chemistry in history

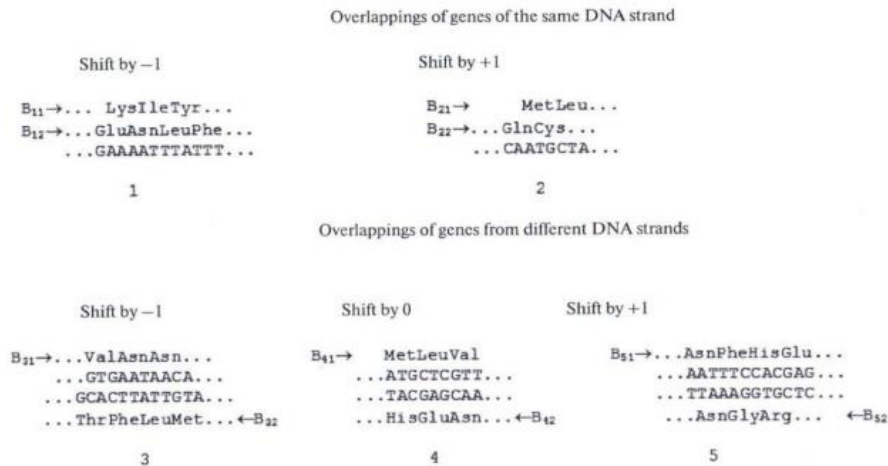
The initial experimental material, at first glance, so strongly reminded those complex discrete systems that had to be dealt with for many years. Due to the fact that such data has already been published in hundreds of articles (and this is only at the beginning of the 90s), I was interested in one question. Where in the world are works containing common positions, principles that describe such a phenomenon, where, after all, is an approach that allows you to work with gene overlaps that differ in just one position or how many such changes are potentially acceptable in a given overlap. It became clear that the multitude of genetic overlaps had to be sorted out in some way, but at first it was not entirely clear how to do this. There was a feeling that this is a lengthy work that requires new approaches that need to be developed based on the indicated experimental data.

After the completion of the first stage of ongoing research, a monograph [6] was published in 2010, which was in great demand (even hackers downloaded it for free more than 70 thousand times before the ban was introduced), and the publisher recently offered me a contract to extend the publication for the next 20 years.

## 2. Introduction

We have carried out a mathematical study of the secrets of the genetic code, starting with the phenomenon of overlapping genes. At present, it is believed that overlapping genes represent, although unusual, but still quite common element of the organization of the genome. Multiple genetic overlaps were found in the deciphered human genome [7] - there were about 1700 of them. The accumulated extensive material on genetic overlaps put forward the task of their thorough and comprehensive analysis. Let us dwell on some results obtained by us on the basis of mathematical analysis. It can be seen that there are only 5 different cases of gene overlaps allowed by the DNA structure (Figure 1), of which the first two refer to overlaps of genes from the same DNA strand, and the remaining three to overlaps of genes taken from different DNA strands.

In Figure 1. only small fragments of real overlaps are presented, and the total length of some of them reaches almost 1300 nucleotides. In addition, the total extent of overlaps can reach more than half of the genome size (GSHV virus [8]).



**Figure 1.** Five possible cases of gene overlaps corresponding to one (1,2) or two DNA strands (3-5). Reading texts in this case is carried out in different directions (indicated by an arrow): from left to right for B<sub>11</sub>, B<sub>12</sub>, B<sub>21</sub>, B<sub>22</sub>, B<sub>31</sub>, B<sub>41</sub>, B<sub>51</sub> and from right to left for B<sub>32</sub>, B<sub>42</sub>, B<sub>52</sub>. These fragments contain only DNA pairs: CG and AT

### 3. Potential possibilities of the code for building overlaps

Next, we set the task of what is the potential of the genetic code to create all these cases of overlap. The answer turned out to be the following [9] - phenomenal potential! Only 16 pairs of amino acids out of a possible 400 can create obstacles to the construction of all 5 cases of overlap. These are the amino acid pairs for just three cases of overlap:

in case 2 it's 5 pairs:

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

in case 3 it's 6 pairs:

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

in case 5 it's 5 pairs:

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

Thus, we have established the integral characteristic of the genetic code, which we denote by  $p$  and which is equal to 16 for the standard code:

$p = 16$  (4)

This result was first published in Doklady Mathematics and was sent to the Nobel Prize winner De Duve (his answer is in the appendix file) after his report at the anniversary celebration in honor of the 300th anniversary of St. Petersburg.

**Table 1.** 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	TGA(ter)@Trp, ATA(Ile)@Met, AGX(Arg)@ter	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 Flatworm 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 Flatworm 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, TCA(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>

#### 4. General property of known natural codes

The result obtained led to the formulation of new tasks. What is the value of  $p$  for non-standard (deviant) codes, the number of which continues to grow? Note that the first non-standard code was discovered in 1979 in a human cell in a separate organelle - in mitochondria: the mitochondrial DNA genes - mtDNA were written with such a code [10]. Only 4 codons were rethought.

Calculations showed that the  $p$  value for all 14 deviant codes (see Table 1) does not exceed the value of 22, or about 5% of the total number of amino acid pairs. In this case, all the same three cases of overlaps were forbidden, as for the standard code, and in addition, one code with a zero value of  $p$  was found. Thus, all natural genetic codes have a small number of prohibitions on the construction of genetic overlaps. This can be seen as a general property of natural codes. The question arises: why do natural codes correspond to this, while the number of gene records with overlaps is immeasurably less than ordinary non-overlapping genes, and what is the role of rethought codons?

#### 5. On the role of rethought codons

Consider now the question of the role of rethought codons. We raised the question of a possible relationship between overlap restrictions (1)-(3) and code variability observed in a number of organisms. 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 Fig. 2), natural rethinking of codons leads to the possibility of constructing genetic overlaps that are forbidden for standard code [11,12].

```

1 (Human)
ATφasa6→ MetAsn...
URF A6L→...Trp...
...ATGAA...
8528

2 (Drosophila yakuba)
ATφasa6→ MetMet...
URF A6L→...Trp...
...ATGATG...
4067

3 (Paracentrotus lividus)
ATφasa6→ MetThrMetThr...
ATφasa8→...TrpGlnTrp...
...ATGACAATGAC...
8680

4 (Apis mellifera ligustica)
ATφasa6→ MetLys...
ATφasa8→...Trp...
...ATGAA...
4585

```

**Figure 2.** Fragments of genetic overlap found in the mitochondria of four organisms whose genes are written in codes that deviate from the standard code. These are overlaps in one strand of DNA. Fragments and names of proteins are given according to publications [10, 13–15]. The number below the nucleotide indicates its number in the genome

Each of the four fragments shows the role of the same permutation: TGA (ter) Trp. This natural permutation is observed for three deviant codes, which correspond to the above fragments, respectively; the second and fourth fragments are written in the same deviant code. Moreover, this permutation is present in all three deviant codes. It turned out that such a permutation makes possible overlaps for pairs of MetAsn (Figure 2.1, this case corresponds to human mitochondrial DNA), MetMet (Figure 2.2), double MetThr (Figure 2.3) and MetLys (Figure 2.4), which are forbidden for standard code. The indicated nucleotide pairs and reinterpreted codons are highlighted. Thus, the size of genomes is reduced due to the possibility of constructing gene overlaps that are impossible for a standard code. Such a contraction for a living cell can be quite large, because the number of mitochondria is usually more than 1 and can reach a million

#### 6. Conclusion

In the monograph [16], we read “The code seems to have been selected arbitrarily”, as well as “Rethinking codons indicate that random changes can occur in the genetic code of mitochondria”.

Our results lead to the conclusion that deviations of the code from the standard carry a completely clear functional load. It follows from the above results that it is possible for all semantic families of codons to write two protein sequences almost unhindered by the same DNA region, and for this the most favorable (according to the combination of amino acids in the overlap) one of the 5 variants of such a compact gene recording can be used (5 cases of overlap). A categorical ban exists for no more than about 5% of amino acid pairs, both for the standard code and for all 14 known non-standard codes. Those. 15 code tables satisfy the same general property. This leaves no chance for any arbitrariness, chance.

In the course of these studies, a mathematical theory of the genetic code was developed, the first part of which was published in Russian [6]. In this regard, the English translation of the annotation and the table of contents of the monograph is given in the Appendix.

## References

- [1] Timur Eneev, Nicolay Kozlov. The Dynamics of Planet Formation. Theory and Computer Simulation. Saarbrucken, Deutschland, LAP LAMBERT Academic Publishing, 2016, 140 p. ISBN: 978-3-659-85958-8
- [2] Kozlov N.N., Kugushev E.I. Algorithm for channel tracing of two-layer LSI // Sb. "Programming of applied systems". - M.: Nauka, 1992. S. 27-32.
- [3] Kozlov, N.N., Kugushev, E.I. Computer simulation tRNA secondary structure folding // Computer Application Biosciences. – 1993. – V. 9, № 3. – P. 253-258. doi: 10.1093/bioinformatics/9.3.253.
- [4] Barrell B.G., Air G.M. and Hutchison C.A. III. Overlapping genes in bacteriophage ΦX174 // Nature. – 1976. – V. 264. – P. 34-41.
- [5] Sanger F., Coulson A.R., Friedmann T., Air G.M., Barrell B.G., Brown N.L., Fiddes J. C., Hutchison C.A., III, Slocombe P.M., Smith M. The Nucleotide Sequence of Bacteriophage ΦX174 // J. Mol. Biol. – 1978. – V.125. – P. 225 -246. Doi: 10.1016/0022-2836(78)90346-7
- [6] N.N. Kozlov MATHEMATICAL ANALYSIS GENETIC CODE MOSCOW, 2010
- [7] Nakayama T., Asai S., Takahashi Y., et al. Overlapping of Genes in the Human. Genome // IJBS. — 2007. — V. 3. — № 1. — P. 14–19.
- [8] Seeger, C., Ganem, D., Varmus, H.E. Nucleotide Sequence of an Infectious Molecularly Cloned Genome of Ground Squirrel Hepatitis Virus // J. Virol. – 1984. – V.51. – P. 367-375. Doi: 10.1016/0166-0934(84)90072-7.
- [9] Kozlov, N.N. A Theorem on the Genetic Code. Doklady Mathematics, **65**, No. 1, 83-87 (2002).
- [10] 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. Sequence and organization of the human mitochondrial genome // Nature. – 1981. –V. 290. – P. 457-464. Doi: 10.1038/290457a0.
- [11] Clary D.O., Wolstenholme D.R. The Mitochondrial DNA Molecule of *Drosophila yakuba*: Nucleotide Sequence, Gene Organization, and Genetic Code // J. Mol. Evol. – 1985. –V. 22. – P. 252-271. Doi: 10.1007/BF02099755.
- [12] Kozlov, N.N. Overlapping Genes and Variability of the Genetic Code. Dokl Biol Sci. Nov-Dec 2000; **375**:677-80. Doi: 10.1023/a:1026631030516.
- [13] Kozlov, N.N. Mathematical analysis of the deviance of the genetic code. Doklady Mathematics 2007, Vol.78, No.3. pp. 851-855.
- [14] Cantatore P., Roberti M., Rainaldi G., Gadaleta M.N. Saccone C. The Complete Nucleotide Sequence, Gene Organization, and Genetic Code of the Mitochondrial Genome of *Paracentrotus lividus* // The J. Biological Chemistry. – 1989. – V. 264, No. 19. – P. 10965 - 10975. Doi: 10.1016/0167-4838(89)90257-4.
- [15] Crozier R.H., Crozier Y.C. The Mitochondrial Genome of the Honeybee *Apis mellifera*: Complete Sequence and Genome Organization // Genetics. – 1993. – V.133 – P.97 – 117. DOI: 10.1007/BF02424475.
- [16] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J., Molecular Biology of the Cell. //- New Jork, London: Gorland Publishing, Inc., 1994. - 1294p.

## Appendix

### Appendix 1. Application

N.N. Kozlov  
MATHEMATICAL ANALYSIS GENETIC CODE  
MOSCOW, 2010

#### Annotation

Based on the study of genes, new properties of the genetic code were established and its most important integral characteristics were calculated. Two groups of such characteristics were distinguished. The first group refers to integral characteristics for DNA regions where genes overlap in pairs, and all 5 cases of overlap allowed by the DNA structure were studied. The second group of characteristics refers to the most extended regions of DNA in which there are no genetic overlaps. The interrelation of the established integral characteristics in the named groups is established. Taking into account the conducted studies, a currently known set containing more than 200,000 genes for 12 large genomes, including 25 613 genes of the human genome, was analyzed. As a result, a number of previously unknown effects were discovered. A comparative analysis of all genes from one cell, but recorded with different natural codes, was also carried out. It was possible to establish two functions in which all rethought codons in the mitochondrial genetic codes (of humans and other organisms) participate, as well as a significant difference in the integral characteristics of such codes compared to the standard code.

#### Table of contents

Preface	page 5
Chapter 1	Introduction 24
1.1.	Genes and proteins
1.2.	Genetic code

### 1.3 Overlapping genes

#### Chapter 2 Mathematical Analysis of Overlapping Genes

##### 2.1. Theorem for overlapping genes

##### 2.2 Proof of the theorem

##### 2.3 Silent mutations in the GSHV genome

##### 2.4. Overlapping genes and irregularities in the genetic code

##### 2.5. Terminator codons in genetic overlaps

##### Conclusion

#### Chapter 3

##### 3.1. On the demand for each of the 64 codons in genetic overlaps

##### 3.2. On the full set of overlapping genes: the case of a shift between genes -1 nucleotide

##### 3.3. About the full set of overlapping genes: the case of a shift between genes +1 nucleotide

##### 3.4. Overlapping genes and variation in the genetic code

##### Conclusion

#### Chapter 4

##### 4.1. Sets Generated by the Genetic Code

##### 4.2. Theorem for the genetic code

##### 4.3. Functional role of rethought codons

##### 4.4. On unusual cases of genetic overlap

##### Conclusion

#### Chapter 5. Integral characteristics of a number of genetic codes

##### 5.1 Hypothetical codes

##### 5.2. Property of all known natural codes

##### 5.3. Two conclusions

##### Conclusion

#### Chapter 6

##### 6.1. Mathematical analysis of structural genes

##### 6.2. Mathematical analysis of the genetic code deviance 6.3. Integral characteristics of the genetic code

##### 6.4. Some calculated characteristics of large genomes

##### Conclusion

#### Chapter 7

##### 7.1. Secondary structure of messenger RNA

##### 7.2. Refinement of the problem statement

##### 7.3. Results of numerical calculations for the secondary structure of MS2 mRNA

##### 7.4. One Feature of Elementary Overlap Sets and Secondary Structure of Messenger RNAs

##### Conclusion

##### Some results

##### Literature

##### Application

### **Appendix 2. The answer from Nobel Prize winner De Duve**

**ICP****Christian de Duve Institute  
of Cellular Pathology**

Under the High Patronage of H.M. Queen Fabiola

Dr N.N. KOZLOV  
Keldysh Institute of Applied  
Mathematics  
Russian Academy of Sciences  
Miusskaya Sq. 4  
125047 MOSCOW  
RUSSIA

Brussels, August 22, 2003.

Dear Dr Kozlov,

Thank you for sending me your interesting paper. There is now good evidence that the genetic code has been optimized by natural selection (see, for example, the article by G. Vogels et al, in Science, vol. 281, pp. 329-331, 1998).

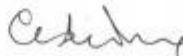
However, the criterion of optimization lies in the minimization of the adverse consequences of point mutations.

If I understand your paper, you propose as criterion maximization of overlapping possibility. This is an interesting point. But is there any evidence that this property has been exploited? Or, in other words, how frequent are overlapping genes in present-day organisms? I am not competent enough to answer that question.

Thank you again for sharing your interesting work.

With best personal regards,

yours sincerely,



C. de Duve

Doklady Akademii Nauk, Vol. 382, No. 5, 2002, pp. 581-592.  
Original Russian Text Copyright © 2002 by Kozlov.  
English Translation Copyright © 2002 by MAIK "Nauka/Interperiodica" (Russia).

---

---

MATHEMATICS

---

---

## A Theorem on the Genetic Code

N. N. Kozlov

Presented by Academician T.M. Eneev October 31, 2001

Received November 8, 2001