Proc. of the Fifth Intl. Conf. Advances in Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2017

ISBN: 978-1-63248-130-6 doi: 10.15224/ 978-1-63248-130-6-05

Computation of the genetic code

Nicolay N. Kozlov, Olga N. Kozlova

***Abstract*—One of the problems in the development of mathematical theory of the genetic code (summary is presented in [1], the detailed -to [2]) is the problem of the calculation of the genetic code. Similar problems in the world is unknown and could be delivered only in the 21st century. One approach to solving this problem is devoted to this work. For the first time provides a detailed description of the method of calculation of the genetic code, the idea of which was first published earlier [3]), and the choice of one of the most important sets for the calculation was based on an article [4]. Such a set of amino acid corresponds to a complete set of re.o.resentations of the plurality of overlapping triple gene belonging to the same DNA strand. A se.o.arate issue was the initial point, triggering an iterative search process all codes submitted by the initial data. Mathematical analysis has shown that the said set contains some ambiguities, which have been founded because of our proposed compressed re.o.resentation of the set. As a result, the developed method of calculation was limited to the two main stages of research, where the fir f st stage only the of the area were used in the calculations. The proposed approach will significantly reduce the amount of computations at each ste.o. in this complex discrete structure.**

Keywords— genetic code, overlapping genes uniform overlap, computation code

Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscowe

I. Introduction

First, con First of all, the question arose: what sets can be used in this case. Our approach was based on the search for all codes satisfying the set of amino acids that record overlapping genes. The author has studied the subject matter of the mathematical analysis of such genes for a number of years [2], and it was this analysis that led to the formulation of this problem.

From the very beginning it was clear that the discussion would mainly focus on the iterative process. Since overlapping of genes contains two, three, etc., amino acids, right up to 6, the first task was to find those overlaps where the same amino acid participates (in overlaps, it participates with different encodings, according to the genetic code) . On this occasion, a special study was conducted.

II. Theorem for homogeneous overlaps

We consider unusual ways of recording genetic information - overlapping genes, when the same DNA portion corresponds to more than one protein. We investigated all 5 possible cases of overlapping of genes resolved by DNA structure, which were studied earlier [5]. This study was based on a mathematical analysis of all 5 possible overlap cases and relied on sets of so-called elementary genetic overlaps-e.o., or overlaps corresponding to a pair of single amino acids. In [6] a brief analysis of such sets is presented, and the final version in [2]. In Fig. 1. A description of the structure of the sets W1-W5 is presented, and are presented by the 4th e.o.. In each of these sets.

The principal position of this research is indicated in [2], where it was shown that the presented list of elementary overlaps can cost any (!) Allowed by the structure of the genetic code, overlapping not only 2 but also all admissible overlap from 3 to 6 Genes. The urgency of the problems is due to the current situation: overlapping genes common in viruses, mitochondria, bacteria and plasmids were found is in eukaryotic of large genomes, including humans, with the

. Description of the structure of sets W1-W5.see fig 4/1 {2 analysis of a set of elementary overlaps for 3 genes overlapping in the same DNA chain. It is established that there are only 307 such overlaps. On the basis of these overlaps, a new problem was posed, connected with the calculation of the genetic code by mathematical methods [11, 12]. The question of why exactly such a set was chosen to calculate the code was based on a theorem that was published relatively recently [4]. We are talking about the calculation and analysis of all homogeneous e.o.ochs. From 2 to 6 genes. Are e.o. which correspond to the same amino acid. Its solution is given by the following theorem.

We call an e.o. -elementary overlap for i amino acids, where, i∈ (2,6). Thus, the e.o. introduced earlier in [2]. Will be referred to as e.o.-2. Figure 2 shows the general re.o.resentation for e.o.-6.

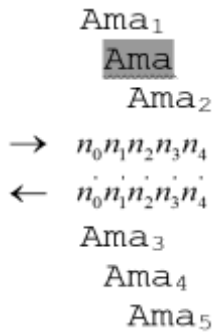

Fig.2. General presentation for e.o.-6. (See the text)

- 5 e.o.-2 for overlaps from one DNA chain:

For the amino acid Ama (isolated by hatching) encoded by the triplet n1n2n3, there are 5 alternative amino acids Ama_1- Ama_5, the encodings of which are formed by -1, +1 shifts in the same DNA chain (→) and -1, 0, +1 in the complementary DNA strand (←). The designations n_i, i∈

Proc. of the Fifth Intl. Conf. Advances in Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2017

ISBN: 978-1-63248-130-6 doi: 10.15224/ 978-1-63248-130-6-05

(0,4) are the nucleotides from the set A, T, C, G; N n'_i i∈ (0,4) - complementary components: i.e. For $n'_{I\ i}$=A; n_i = T ; n_i=C, n'_i =G for any i∈ (0,4) and vice versa. In order to sequentially isolate e.o.-2 for all 5 cases of pair overlaps from [2,3], in Fig. 3 one should consistently leave one of the 5 pairs of amino acids: (Ama, Ama_i), i∈ (1 ,5). In order to sequentially isolate all overlap cases for e.o.-3 in Fig. 3, it is necessary to leave Ama and a pair of amino acids out of 5 possible ones or we have 10 cases of triple overlap, etc. We turn to the search for a homogeneous subset for all possible e.o.-2- -6, or subsets, in each e.o.. Which is the same amino acid. We denote it by $W_{2\text{-}6}$. To represent homogeneous elementary overlaps, the following notation is used: Y: T, C; X: A, G; N: A, G, C, T.

,

First of all, it is necessary to exclude from consideration all homogeneous overlaps in which two strands of DNA participate. Consideration of these overlaps requires the introduction of a double strand of DNA - this is an additional condition in the problem. Eliminating such homogeneous overlaps, we proceed from the principle of constructing an algorithm with a minimum number of conditions. Therefore, in our examination there remain only homogeneous overlaps belonging to the same DNA chain: for pairs of amino acids (1), there are only 5 of them and similar overlaps for three amino acids (3) -th total of 4. Thus, we selected the main working sets E.o., namely, those in which these homogeneous overlaps are present. The final version of these sets is presented on pages 312-319 in [2].

.ontains two nucleotides n1n2; The codon n2n3n4 for Ama2 overlaps with codon n1n2n3 for Ama - the overlap contains two nucleotides n2n3. As a result, the triple overlap contains only one common position n2.

B. Elements of sets of combinations of amino acids, formed on the basis of elementary overlap from Fig. 3A. On the left there is one element of the set U1, in the center there are two elements of the set U2 and to the right one element of the set U3.

.

Amino acids shown in Table 1. Elements are given in the view that is used in this task. Each of the elements consists of three lines: upper, middle and lower. The named amino acids Met and Arg are shown in the middle line.

Formulation of the problem. Introduction of a compressed set.(see309-319 in {2{(

H/

Fig. 5 The elements with numbers 1-8 and 279-307 from the set U3, corresponding to the first (Met) and last (Arg) a of the main set-U3: for each Ama of this set (it is indicated in the corresponding cell) is given the Ama1 amino acid along the abscissa axis, and on the ordinate axis - Ama2 (see Fig. 1A). It turned out that the resulting re.o.resentation is not homogeneous, but contains multiple ambiguities: these are the cases when more than one Ama value corresponds to the same Ama1 and Ama2 values - from 2 to 4. These cases are shaded in this figure, they are denoted by A1-A13, i.e. There are only 13 of them, although the figure shows 34 hatchings. The fact is that in this figure A6, A9 and A10 are re.o.resented 4 times, - A1, A5, mino acids from the list

Fig.6. The compressed re.o.resentation for 307 elements A7, A8 and A11 - three times, A3 and A4 - 2 times, and A2, A12 and A13 - only by One time. We have A1 Gln, Leu; A2: Gln, Val, Leu; ... A13: Gln, Pro, Ala.

III. Formulation of the problem

Let's have a set of 4 letters: N: a, b, c, d, and also triplets - any triples of these letters, there are 64 in all. Moreover, each of the 20 canonical amino acids can be encoded by an arbitrary combination of such triplets. The task is to search for all the genetic codes that correspond to all the elements designated above the three sets U1, U2, U3, corresponding to the genetic experiments.

For the future, we use standard three-letter abbreviations for each of the 20 amino acids.

In [11] we talked about ste.o.s to calculate this problem, but did not say how the required elements were selected at the ste.o.. For such a selection, a special re.o.resentation of the basic working set was introduced. This result is published in detail for the first time.

We introduce one concise re.o.resentation for 307 elements of the principal set-U3. In Fig. 6, for each Ama of this set (it is indicated in the corresponding cell), the amino acid Ama1 is plotted along the abscissa axis, and the ordinate is Ama2 (see Fig. 3A). It turned out that the resulting re.o.resentation is not homogeneous, but contains multiple ambiguities: these are cases when more than one Ama value corresponds to the same Ama1 and Ama2 values. It turned out that the number of ambiguities in the range from 2 to 4. All these cases are shaded in Fig. 6 and they are denoted by A1 - A13, and their decoding is given in the caption to this figure.

It should be noted that these ambiguities correspond to the values of Ser, Leu, Arg, both along the abscissa axis and along the ordinate axis. However, the most significant area in Fig. 6, which corresponds to the cases where on both axes there is none of the amino acids from the Ser, Leu, Arg. For our calculations, the last region is reduced, eliminating from it all cells containing Ser, Leu, Arg. In Figure 7, the shading corresponds to the three amino acids mentioned, and the non-zero elements of the unshaded region have the following property: each Ama value is unique for the corresponding pair Ama1 and Ama2.

The above property allowed us to refer to the first stage of the calculation, when the calculation of the encodings for all elements is made Ama value on the basis of the encodings for the corresponding pair Ama1 and Ama2. The results of the ste.o.-by-ste.o. solution of the problem are presented in Table 2, but the most important stage of the study was the question of finding the initial approximation.

IV. Solution of the problem

We use the standard three-letter abbreviations for each of the 20 amino acids listed in the first column of Table 1. We have a set

A0: Amai, i (1,20). (5)

We introduce the definition. Let us turn to the previously introduced homogeneous overlaps. As before, we call a combination of amino acids, constructed on the basis of an elementary genetic overlap, homogeneous if the same amino

ISBN: 978-1-63248-130-6 doi: 10.15224/ 978-1-63248-130-6-05

acid participates in it. For homogeneous elements of the set we have

Property. Let the encodings Ama for homogeneous u3 have one of the following three re.o.resentations:

$$n1N1N2\ ,\qquad N3n2N4\ ,\qquad N5N6n3\ , \qquad (6)$$

¬

Where small letters denote the unit components of the set N, and large - some subsets of this set, up to N. Then homogeneous u3 can exist only if at least one base triplet or triplet with three identical letters is used.

For the proof we successively substitute each of the re.o.resentations (6) in u3:

Ama Ama Ama

Ama Ama Ama (7)

Ama Ama Ama

н.п. n1n1n1n'1n'2 n'3n2n2n2n'4 n'5n'6n3n3n3

.

3		Met	Trp	Phe	Tyr	His	Asn	Asp	Cys	Gln	Lys	Glu	Ile	Val	Pro	Thr	Ala	Gly
	Met													Cys			Cys	Trp
	Trp							Gly				Gly		Gly			Gly	Gly
	Phe			Phe														
	Tyr		Met	Ile	Ile	Thr			Met	Thr					Thr			
	His		Met	Ile	Ile	Thr			Met	Thr					Thr			
	Asn		Met	Ile	Ile	Thr			Met	Thr					Thr			
	Asp		Met	Ile	Ile	Thr			Met	Thr					Thr			
	Cys		Val	Val	Val	Ala			Val	Ala					Ala			
	Gln	Asn					Lys				Lys		Asn			Asn		
	Lys	Asn					Lys				Lys		Asn			Asn		
	Glu	Asn					Lys				Lys		Asn			Asn		
	Ile	Tyr		Phe									Tyr			Tyr		
	Val	Tyr		Phe									Tyr	Cys		Tyr	Cys	Trp
	Pro	His				Pro	Gln			Pro	**Gln**		His		Pro	His		
	Thr	His				Pro	Gln			Pro	Gln		His		Pro	His		
	Ala	His				Pro	Gln			Pro	Gln		His		Pro	His		
	Gly	Asp	Val	**Val**	Val	Ala	Glu	Gly	Val	Ala	**Glu**	Gly	Asp	Gly	**Ala**	Asp	Gly	Gly

Fig.8.

Where n'i is the single component of the set Ni, where i∈ (1,6), and the string na. - nucleotide sequences that are formed after this substitution. In the first case, in (3), the base codon n1n1n1 was used for encoding amino acid Ama from the bottom position, in the second - n2n2n2 - from the middle position, and in the third - n3n3n3 - from the top position.

ISBN: 978-1-63248-130-6 doi: 10.15224/ 978-1-63248-130-6-05

We turn to homogeneous u3 from the set U3, which turned out to be 4:

Lys Phe Pro Gly
Lys Phe Pro Gly (8)
Lys Phe Pro Gly

Within the framework of the assumption specified in the Property, the following ste.o.-by-ste.o. process of searching for a genetic code is proposed; See Table 2. Ste.o. 1. Amino acids from (8) will assign the corresponding base codons. This assignment is not unique. However, in our approach, the set of letters N: a, b, c, d is not correlated with the canonical set of 4 nucleotides; This will be discussed at the end of the paper. Therefore, we will continue to operate with only one of the re.o.resentations for the amino acids from (4), which we assign respectively the following basic triplets:

Lys: aaa, Phe: bbb, Pro: ccc, Gly: ddd (9)

For further calculations, we turn to some generalized data on the sets U2 and U1, which are given in Table 1.

Step. 2. From Table 1 it follows that, as in column m12 (the number and the list of overlapping amino acids on 1 and 2 bases are indicated), and in column m23 (similar data for 2 and 3 bases) do not contain mutual overlap between amino acids from (8). Such overlaps take place only one position and belong to the set U1. We have:

Lys Phe Pro Gly
Lys Phe Lys Lys
(Gly) (Pro) (Pro) (Phe)
(Gly) (Pro)
(Phe) (Gly)
(10)

Lys Phe Pro Gly
Lys Phe Pro Gly
(Pro) (Pro) (Phe) (Lys)
(Gly) (Gly) (Gly) (Pro)

Where the first 4 elements of u1 correspond to overlaps for 3 positions (in parentheses the alternative variants are indicated, see column m3-1 from Table 1), the next 4 - overlaps for the first positions (see column m1-3). The formal substitution of the base codons from (9) into (10) leads the encodings of all 4 amino acids to the fact that they become ambiguous in the 1 st and 3 rd positions. For the sake of clarity, we present the derivation of just two amino acids from (9) -Lys and associated with it according to the first overlap of (10) -Gly. According to the above, we have: Lys should be encoded by a set of triplets X1aX, and Gly - X1dN, where X1: a, d, c; X: a, d. Then there is an overlap of Lys with Gly in two positions, which is impossible according to Table 1. It also does not allow two other possibilities: Lys can not be encoded by a set of triplets X1aa if Gly is ddN, and also Lys can not be encoded by a set of triples aaX if Gly is X1dd. There are still two possibilities: both Lys and Gly are coded by ambiguous codons for the same positions. The case of Lys: X1aa, Gly: X1dd encoding is impossible, as the condition in Table 1 can not be satisfied: the number m23 for Gly is 5 according to Table 1. (And for a similar encoding Gly there can be a maximum of -4). Therefore, there remains the only possible option for the encodings in question, when there are ambiguities in the third position. Similarly, you can set the encoding for the remaining Phe and Pro pairs. In the end, we get:

Lys: aaX, Phe: bbY, Pro: ccN, Gly: ddN, (11)

From Table 1 it follows that the value of m12 does not exceed the number 4 for the amino acids from (5) with the numbers from 1 to 17. The number 4 means that the first and second positions can be single-valued, which can not be said for m12 for Ser, Leu, Arg, for which these values are 7.8.7, respectively. Therefore, in the next ste.o.s 3-7 only amino acids from (5) with numbers up to 17 will be considered. Note that the calculation of the encodings for all amino acids from (5) is carried out according to the method published in [3]

Step 3-by-step 7 calculation in the uniqueness domain

The solution search in step 3 is illustrated in Fig. 8 (step. 3), where the reduced reduced unambiguity region is presented. For each of the four amino acids from (9) we carry out two bands (gray hatching in Fig.) For Ama1 (horizontal strip, figure 3 outside the figure indicates the step. number) and Ama2 (vertical strip). As a result, at the intersections of these bands we find only 4 amino acids: Gln, Glu, Val, Ala, which are shown in the figure as bold. Taking into account the acce.o.ted standard record and the codings from (9), we have

Lys Lys Phe Pro
Gln Glu Val Ala
Pro Gly Gly Gly (12)
ccaa ddaa ddbb ddcc

Where n.s. is the nucleotide sequence. From (12) we have single-valued encodings for 4 amino acids: Gln, Glu, Val, Ala, and with (9) we find:

Gln:caX, Glu:daX, Val:dbN, Ala:dcN, (13)

Thus, it is shown that the use of the introduced reduced set leads to minimal costs in the ste.o., compared to a direct search for 307 elements of the main set - U3.

Step 4-Steh 7-on analogu

Proc. of the Fifth Intl. Conf. Advances in Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2017

ISBN: 978-1-63248-130-6 doi: 10.15224/ 978-1-63248-130-6-05

Search for a solution in the field of ambiguity

Step 8. Finding solutions in an area where the values of Ama1 and Ama2 belong not only to these 17 amino acids. On the basis of Fig.7 We get

Gln Pro Ala Val
Ser Ser Ser Ser
Phe Ile Gln Gln

Tyr Phe Trp Tyr Trp
Leu Leu Leu Leu Leu
Thr Ala Pro Val Val (22)

Glu Val Ala Gly Gly Gly
Arg Arg Arg Arg Arg Arg
Lys Ala Ala Pro Thr Lys

From these overlappings we find the following encodings:

Ser: bca,bcc,adY; Leu: cbX,cbb,bbX; Arg: cdN,adX. (23)

Step 9. From the ambiguity region in Fig. 2, we select cases containing two amino acids from the set Ser, Leu, Arg and giving solutions different from (23). We have:

Arg Leu Ser
Ser Ser Leu
Val Ile Pro (24)

From the first and second overlap we find: Ser: bcd, bcb, and from the third - Leu: cbc. The final encodings for Ser, Leu, Arg are presented in Table 2. And the total number of semantic triplets for all 20 amino acids from in this table is 61. An additional check shows that all elements of the ambiguity region do not contain any other solutions. Three triplets: baX, bda are not defined when using any elements of the sets U1, U2, U3; They supplement the total number of triplets to 64.

When passing from the set of letters a, b, c, d to the canonical nucleotides A, C, T, G, 24 similar genetic codes can be obtained. Only one of them is standard, with a = A, b = T, c = C, d = G, and triplets baX, bda become TAA, TAG, TGA; They play a role

Acknowledgment

The work was supported by Russian Foundation for Basic Research (project codes 16-01-00018,17-01-00053)

	1	2	3	4	5	6	7	8	9
Met							abd		
Trp				bdd					
Phe	bbb	bbY							
Tyr						baY			
His						caY			
Asn						aaY			
Asp						daY			
Cys				bdY					
Gln			caX						
Lys	aaa	aaX							
Glu			daX						
Ile							abM		
Val			dbN						
Pro	ccc	ccN							
Thr					acN				
Ala			dcN						
Gly	ddd	ddN							
Ser								bca bcc ddY	bcd bcb
Leu								cbX cbb bbX	cbc
Arg								cdN adX	
Σ	4	12	24	27	31	35	43	58	61

Table 1. Steps 1-9 of the iterative process of calculating the genetic code. The last line indicates the number of calculated codons after this step.

Notation: X: a, d; Y: b, c; M: a, b, c; N: a, b, c, d.

Fig.7. The who translated this text from Russian. The work was supported by Russian Foundation for reduced region of Fig. 6: there are areas in which the code is calculated. All the shaded regions are cut off for the reasons indicated above, and the main area in the calculation is re.o.resented without shading ……

.References

[1] N.N. Kozlov, T.M. Eneev: Fundamentals of a Mathematical Theory of Genetic Code. Doklady Mathematics, 2017, Vol. 95, №2, pp. 1-3

[2] N.N. Kozlov. Genetic Code: A Mathematician's Point of View. Palamarium Academic, Hamburg, 2014, 336 p. ISBN: 978-3-639-63268-2

[3] Kozlov, N.N. Method of Computing a Genetic Code. Doklady Mathematics 2015, Volume 91, №3, pp.263-266

[4] Kozlov, N.N. On Overlaps of More than Two Genes: A Theorem for Homogeneous Overlaps. Doklady Mathematics 2015, Volume 91, № 3, pp. 1-3

[5] Kozlov, N.N. A Theorem on the Genetic Code. Doklady Mathematics, **65**, No. 1, 83-87 (2002).

[6] Kozlov, N.N. Sets generated by genetic code. Doklady Mathematics 2008, Vol.78, No.3. pp. 851-855.

[7] Nakayama T., Asai S., Takahashi Y., et al., Overlapping of genes in the human. Genome NJBS 2007, vol.3, no. 1, p. 14-19.

[8] Kozlov, N.N. Ambiguity in sets generated by the genetic code. Doklady Mathematics 2010, Volume 81, No.3 pp 364-367

[9] Kozlov, N.N. One function of ambiguities from the sets generated by the genetic code Mathematical Models and Computer Simulations January 2013, Volume 5, Issue 1, pp 17-24.

[10] Kozlov N.N. Three Functions of Ambiguities Generated by the Genetic Code. Mathematical Models and Computer Simulations 2015, Volume 7, Issue 5, pp 401-408.

[11] Kozlov, N.N . Computation of the genetic code. Doklady Mathematics 2010, Volume 82, No.3 pp 535-539

[12] Kozlov, N.N Computation of the genetic code Mathematical Models and Computer Simulations February 2012, Volume 4, Issue 1, pp 36-46