

To FUNDAMENTAL BASES of DISCRETE MOLECULAR BIOLOGY

Eingorin. M. Ya.

Nizhniy Novgorod State University, SRDE "SCIT", p. o. 35, Nizhni Novgorod - 136, Russia, 603136. E-mail: skit@unn.ac.ru

It would be difficult to overstate the importance of the adequate understanding of gene texts coding principles, which exist in nature. The objective of my research is to establish gene texts coding fundamentals. The paper is based on the elementary facts of biology: four complementary nucleotides organized in pairs form DNA; DNA is essentially a double spiral with orthogonal-directed strands; there exist closing codons and an opening codon with its promotive zone; there are 20 amino acids; we now know various dialects and gene degeneracy, transformations and correction phenomena; etc. I have also made use of multiple-valued logic and combinatorial mathematics in laying the theoretical footing of the paper. The basis thus provided allows me to build up a codonogram and an aminogram, three-dimensional natural constructions with digital independent variables U, C, A, G. The constructions reveal new patterns and rules of coding in molecular biology fundamentals (elementary codon group or ECG is the basic coding element; there are complementary tubes; amino acids polarization on ECG shows uniformity; etc). New correspondences are established alongside some principles of dialect amino acids – codons tables' compilation. The codonogram secondary coding, which is possible thanks to redundancy, accounts for the possibility of latent coding layers (LCLs) distinguishing in DNA, RNA and mRNA. Coming into contact with enzymes LCLs provide for gene addressing, synchronization, correction and other functional transformations. The paper describes possible mechanism and patterns of variability in nature. It also demonstrates regularity and variability, which underlie coding. Given in the paper are: codonogram, aminogram for universal genetic code (UGC), basic rules for LCL building, references to the sourcebook and an example of decoding, language for further building. Most of the presented material has never been published before. The paper clearly demonstrates that genetic engineering without proper knowledge of basic coding rules existing in molecular biology might lead to irreversible results for humankind. Germs will survive, however.

Codonogram, aminogram, latent coding layers, coding, decoding, gene, DNA, RNA, mRNA.

Since the 40s-60s of the now last century - the time when the famous scholarly works were published - we have not advanced much in understanding genetic texts coding/decoding principles. We have seen a rapid development of genetic engineering but a great many of DNA coding principles have not been discovered so far. This is fraught with danger. Besides, we cannot possibly carry out computer-aided genetic analysis and synthesis without such knowledge. To establish some of the new patterns and principles I have referred to some 30 molecular biology elementary facts, which are the actual basis of this paper.

Laws of biology lay a theoretical footing of the paper, the research instruments being my works on multiple-valued logic and combinatorial mathematics, experience acquired, intuition etc.

As paper [1,2] demonstrates, the codonogram incorporates DNA general coding principles common for all dialects, whereas principles that are individual for each particular dialect are embodied in the aminogram. The present paper serves as continuation of paper [1] and establishes new interdependencies and patterns of coding. It has been my endeavor to disclose natural mechanisms implicit in the dialect codonogram and aminograms and partially in their transformation on DNA. I understand the functional and informational aspect of the genetic constructions as the most significant issue and reveal it as far as possible.

The principle results of the research are as follows:

1. We represent all the 2^6 codons $X_1X_2X_3$ as a cube – codonogram - Q_0 with $X_i = \{t(u), c, a, g\}$, $i = 1,2,3$. X_i can be expressed as:

$$X_i = (x^p x^v)_i = x^p_i x^v_i,$$

x_i^p and x_i^v being the pyrimidine/purine component and the double/triple (V_2/V_3) hydrogen bond component respectively, $x_i^p x_i^v = \{0,1\}$. The table 1 - table of coding nucleotides. We separate 16 elementary codon groups or ECGs, each containing 4 codons, along X_3 direction in Q_0 . The ECGs fall into 8 basic groups or BECGs and 8 alternative groups or AECEGs (see fig. (1-5)).

We combine the ECGs into four «tubes» (6). Connected (\sim) (complementary) nucleotides, codons, ECGs are situated in the tubes diagonally. In this case DNA coherence can be defined as $x_i^p = N x_i^{p'} \equiv \underline{x}_i^{p'}$ and $x_i^v = x_i^{v'}$. Coherent codons of DNA strands (5'-3') and (3'-5') lie in planes

$$(t\sim a)_1 X_2 X_3, X_1(t\sim a)_2 X_3, X_1 X_2 (t\sim a)_3 \text{ or } (c\sim g)_1 X_2 X_3, X_1(c\sim g)_2 X_3, X_1 X_2(c\sim g)_3.$$

And each X_i have: $P_i \sim P_u$ bases and $V_2 \sim V_2$ or $V_3 \sim V_3$ hydrogen bonds.

Nucleotide mass: $m_t = 322$ g/mol, $m_c = 307$ g/mol, $m_a = 331$ g/mol, $m_g = 347$ g/mol.
Coherent nucleotides sum m_i :

$$m_t + m_a = m_{ta} = 653 \text{ g/mol and } m_c + m_g = m_{cg} = 654 \text{ g/mol. } (m_{ta} \cong m_{cg}) = \mathbf{m}_{tg} - \text{DNA constant.}$$

The mass of the coherent codons in the codonogram is $3\mathbf{m}_{tg}$. The mass of the coherent ECGs is $12\mathbf{m}_{tg}$. In this case any DNA results mass-balanced for \mathbf{m}_{tg} is invariable.

In Fig. in Q_0 shows other ties and interdependencies for nucleotides, codons and ECGs.

2. We present R-dialect codons - amino acids (C-A) Table as a system of logical equations (SLE)_R. (SLE)_R correspondences:

$$\underset{pz}{V}(X_1 X_2 X_3)_{pz} \sim A_z$$

For: $1 \leq pz \leq 8$, $z = (\underline{1,21})$, A_z stands for aminoacid (9), $(X_1 X_2 X_3)_{pz}$ for codons and V is the disjunction sign for «pz», $1 \leq p \leq 2^6$.

We «cover» Q_0 with (SLE)_R system. As a result we have Q_R aminogram developed along A_z in X_3 direction (Fig.(18), heavy lines). All the BECGs are «covered» with one A_z and the AECEGs - with two A_z and A_w (10), which have common P or NP polarization (11). In Q_R 8 ECGs are «covered» with P, and 8 ECGs are «covered» with NP A_z . Other correspondences for Q_R are given in the text and Table 2 of the figure, as well as an instance of (SLE)_{UGC} for UGC (Universal Genetic Code). It is assumed that the covering AECEGs in A_{z1} and A_{z2} peptide chains are alternative. Additional «cross» tRNA for $A_{z1} \sim A_{z2}$, whose functioning depends on various conditions, are responsible for proteins probabilistic variability. There occurs certain ECG «covering» deviations for some Q_R .

3. Table 2 shows that DNA analysis and synthesis starts with X_2 , which determines one of the planes $X_1(u,c,a,g)_2 X_3$ and the combinations of its properties as respects B, A, P and NP. X_1 determines the plane ECG with its characteristic properties. $X_3 = x_3^p x_3^v$ determines Pi/Pu and V_2/V_3 of the chosen ECG for DNA, RNA and mRNA.

The beginnings ECG, lying in a plane $X_1 X_2 u$, as have «rough alternative» (RAECG) "covering" aminoacids (According to the Table - 2: BP - 3, ANP - 5, AP - 5, ANP - 3) also can, with the certain reserve, to "replace" each other. On planes X_2 alternative on RAECG is allocated:

$$\begin{aligned} &\text{for } X_2 = t - X_1 = t \sim X_1 = a \text{ and } X_1 = c \sim X_1 = g; \\ &\text{for } X_2 = c - X_1 = t \sim X_1 = a \text{ and } X_1 = c \sim X_1 = g; \\ &\text{for } X_2 = a - X_1 = t \sim X_1 = c \sim X_1 = a \sim X_1 = g. \\ &\text{for } X_2 = g - \text{the alternative on RA - is not present.} \end{aligned}$$

Here we have a 100% redundancy in matching $x^p_1 = \{0,1\}$, $x^v_1 = \{0,1\}$, $x^p_2 = \{0,1\}$, $x^v_2 = \{0,1\}$ on RAECG and $x^p_3 = \{0,1\}$, $x^v_3 = \{0,1\}$ on BECG and AECG. Q_R covering redundancy makes it possible to reveal the existence of latent coding layers or LCLs «connected» with DNA, RNA and mRNA codons. LCLs do not depend on the Q_R and can be disclosed by the additional coding (20) of Q_0 codons on the basis of the logical equations system for coding (LES-C) (Table 2). The Fig. presents the LES - C of three coding layers $|J_a|$, $|J_b|$, $|J_c|$. Thus,

the layer $|J_a|$ is defined X_1X_2 ,
 layer $|J_b| - x^p_3 = \{0,1\}$ and
 layer $|J_c| - x^v_3 = \{0,1\}$.

To «expose» LCL J_i we associate binary codes J_a, J_b, J_c or other codes with the peptide chains. J_i layer is described in terms of the codon code pair, E^1J_i and E^2J_i being the odd and the even members respectively. The pairs $\{E^1J_i, E^2J_i\} \subset EJ_i$ being chosen correctly, the $|EJ_i|$ chain is constituted by a sequence of overlapping ($|EJ_c|$) zones or joined end-to-end non-overlapping ($|EJ_a|, |EJ_b|$) symmetry groups.

Zones and groups have centers, half-groups, half-group fragments and sub-groups, which are symmetric about the center. Symmetry zones and groups are characterized by special patterns and apparently «converse» with enzymes in the language of symmetries. Each LCL of a gene has its own basic symmetry zone.

Laws of coding SLE:

1. On coordinate directions X_1, X_2, X_3
 Are postponed nucleotids in the order: t, c, a, g.
 Everyone nucleotid is submitted by components x^p and x^v , as it is shown in the Table - 1:

Aminogramm UGC.

| Tab -1 | Pi | Pu | V2 | V3 | Cod X+ | Trin nuc. |
|------------------------------------|---------|---------|---------|---------|-----------|-----------|
| $X_1 \downarrow$ Cod \rightarrow | $x^p=0$ | $x^p=1$ | $x^v=0$ | $x^v=1$ | $x^p x^v$ | |
| u (t) (Tim.) | 0 | - | 0 | - | 0 0 | Ur. |
| c | 0 | - | - | 1 | 0 1 | Cit. |
| a | - | 1 | 0 | - | 1 0 | Ad. |
| g | - | 1 | - | 1 | 1 1 | Gu. |

2. Everyone codon $X_1X_2X_3$ on Pic. has situation in axes $X_i = \{t, c, a, g\}$, $i=1,2,3$.

3. Are available 16 ECG of a direction X_3 .

4. Eight base BECG: tc1, gc1, ac1, gc1, ct1, gt1, cg1, gg1, in a Fig. - large points.

5. Turn BECG on 180° along an axis X_3 on x-x Gives eight AECG: ta1, ca1, aa1, ga1, tt1, at1, tg1, ag1, - small points, look item 19.

6. In codonian "skeleton" of item (1-6) are formed four codonian "tubes": I-cc1~gg1, cg1~gc1; II-tc1~ag1, tg1~ac1; III-ct1~ga1, ca1~gt1; IV-tt1~aa1, ta1~at1.

7. Coherent nucleotids and codons are located in tubes on their diagonals.

8. Tub I have only BECG, IV - only AECG, II and III on diagonals - mixed.

9. On aminoacids in Q_R :
 S-ser~R-arg, T-the, Y-ter, D-asp~E-glu, N-asn~K-lis, H-his~Q-gin ← "P", L-lei~F-pfe, P-pro, V-val, A-ala, G-gly, I-ile~M-met, C-cys~W-trp ← "NP"

10. Any equality (SLE)₈ "covers" with aminoacid one BECG at four codons, BECG and part AECG at greater 4 and part AECG at smaller 4. At 8 codons - covers two BECG.

11. On Fig. "P" ECG are designated by circle around of a point-codons; NP - without a circle.

12. All starting codons are located on a plane $X_1X_2X_3 \sim (x^p x^v)(00)_2(1x^v)$.

13. All codons Ter. dialects, except for mitochondrian man, are located on Planes $X_1X_2X_3 \sim (00)_1(x^p x^v)(1x^v)$.

14. The planes on 12 and 13 are perpendicular.

15. All codons for any dialect R "Are covered" aminoacid or Ter.

16. 17 dialects R of the tables C-A are known.

17. Laws of formation aminogram Are correct practically for all known on Today of dialects Q_R .

18. Example (SLE)₈ of a dialect R Code UGC. (SLE)₈ucc.

ttt v ttc ~ F; tta v ttg v ct1 ~ L;
 att v atc v ata ~ I; atg ~ M;
 gt1 ~ V; tc1 v agt v agc ~ S;
 cc1 ~ P; ac1 ~ T; gc1 ~ A;
 tat v tac ~ Y; taa v tag v tga ~ Ter.
 cat v cac ~ H; caa v cag ~ Q;
 aat v aac ~ N; aaa v aag ~ K;
 gat v gac ~ D; gaa v gag ~ E;
 tgt v tgc ~ C; tgg ~ W;
 cg1 v aga v agg ~ R; gg1 ~ G;

19. $X_1X_2t \vee X_1X_2c \vee X_1X_2a \vee X_1X_2g = X_1X_21$

20. Example SLE-C.

tttattvatcatvaatvgatvgtvagt ~ 000
 ctvtgttctvctvactvgtvctvgtvgt ~ 100
 ttavataavaaavaaavaaavaaava ~ 010
 ctavgtavcavccavcavcavcavcavcav ~ 110
 ttcvatcavcavcavcavcavcavcavcavc ~ 001
 ctvgtcvcvccvccvccvccvccvccvccvcc ~ 101
 ttvatgvaagvcaagvaagvcaagvcaagvcaag ~ 011
 ctvgtgvcgvcgvcgvcgvcgvcgvcgvcgvcg ~ 111

Ja Jb Jc

Let us consider the interdependence of DNA LCL spirals (5'-3') and (3'-5'). We can see that J_c for spiral (5'-3') coincide with J_c' for spiral (3'-5'), J_a coincide for tubes I and IV and invert for tubes II and III. They also invert for J_b layer. The (') symbol is used for spiral (3'-5'). Each LCL is

associated with a certain physical/chemical (Pi, Pu, V₁, V₂, P, NP) property of the nucleotides, their combination and position in Q₀. It has its own pattern and provides for DNA and mRNA synchronization, reaction and correction, DNA convolution etc. LCL layers serve as sort of matrices comprising the characteristic properties of a gene and describing the way it reacts with cell enzymes.

The «start» and «stop» codons position on E¹ or E² in the chain changes their function. The figure displays some of the Start (12) and Ter. (13) codons characteristics. As controlling codons they are all situated in the area described as $x^p_3 = 1$. The planes they [3,4] are situated in are perpendicular (14). As appears from the abstract and the paper, Q_R is not degenerate, there are commas and periods in DNA, which in turn appears mass-balanced.

As appears from the poster paper, there exist well-ordered genetic text coding and, consequently, decoding principles, which are determined by the structure of the dialect codonogram and aminograms. Algorithms built on these principles will permit analyzing the existing genetic structures and synthesizing new ones for the benefit of biology, pharmaceuticals and medicine when the correction of genetic abnormalities is necessary. And above all, the dialect codonogram and aminograms, their chemical and equivalent mathematical (logical and geometrical) structural patterns as well as their coherence and interdependencies exist irrespective of DNA, mRNA and other compounds and present an essential set of rules/laws for the formation of the latter. The dialect codonogram and aminograms feature an exceptional regularity and complexity, which gives grounds to assume that they as well as genes and DNA were created artificially and consequently to postulate the artificial origin of life on the Earth.

The following is a brief summary of the cycle of publications:

- 1) As is evident from the three-dimensional codonogram and aminogram, all structures existing in Nature feature regularity implicit in the rules for making up (C-A)_R dialect tables.
- 2) Q₀ codonogram is a basic three-dimensional structure built up for representing codon and nucleotide properties and interdependencies (Pi, Pu; V₂, V₃). The aminograms of Q_R dialects, nucleotide and amino acid parameters are basic three-dimensional structures for amino acids and their properties («P», «NP»; «Б», «А»). The codonogram and aminograms incorporate DNA and mRNA coding principles, according to which new interdependencies and interpretations for facts of molecular biology are established and new interdependency patterns for the structural elements are set up.
- 3) The paper demonstrates that u(t), c, a, g nucleotides have such an arrangement that feature distinct physical and chemical properties along Pi – Pu, V₂ – V₃ and have binary codes that match their properties. Coherence is functionally implicit in V₂ – V₃ nucleotide.
- 4) The mathematical apparatus of the elements of positional two-valued and multi-valued logics as given in the paper describes DNA and RNA coding mechanism and patterns.
- 5) The functional purpose of the parameters of codon nucleotides in their connection with Pi, Pu bases, V₂, V₃ number of hydrogen bonds and P, NP polarizations is formulated.
- 6) The coherent nucleotide, codon and ECG weight invariable is introduced in the codonogram.
- 7) DNA is mass-balanced by coherent nucleotides and codons, the weight invariable being **m_{tg}** and **3m_{tg}** respectively.
- 8) Revealed in the paper is the profound internal structure in Q_R forming, as well as certain correspondences underlying it.
- 9) (SLE)_R and aminogram redundancy is employed in the formation of DNA and mRNA latent coding layers (LCLs).
- 10) The gene carries *two informational sequences* – those of protein synthesis and LCL-control, which interact with the cell enzymes.

- 11) The codonogram allows setting LCL formation principles and elaborating a uniform LCL latent coding system, which is based on the codon and physical and chemical nucleotide properties.
- 12) The principle of uncertainty in LCL coding is formulated.
- 13) Elements of the LCL language are given. Symmetry and patterns of symmetry zones and groups constitute the principal language for LCL coding and DNA and RNA formation.
- 14) LCLs are formed on the basis of Pi, Pu bases combinations and in accordance with the number (V_2, V_3) of the hydrogen bonds of codon nucleotides.
- 15) It is shown, that all DNA and RNA in layers of management through codonogram and aminograms are constructed on the basis of two-valued and two-multiple-valued logic.
- 16) It is assumed that codon division proceeds multiply of LCL groups.
- 17) Each LCL at the gene initial part has a sort of «cap», which includes Start codon and differs in size and pattern from the subsequent periodic structure.
- 18) LCL yields possible principles of synchronization, defects detecting and their correction in a genetic strand.
- 19) As is clear from any of the constructions, the genetic code has «commas» and «periods» owing to LCL; it is not degenerate and has overlapping and non-linear nature.
- 20) The principles of making up (C-A)_{UBC} codon groups and human mitochondria ensure LCL formation optimization. Some principles of dialect aminogram formation appear from the paper as well.
- 21) It is assumed that the variability mechanisms of organisms that exist in nature are implicit in the aminograms (owing to AECG and RAECG coverings).
- 22) It is assumed that the origin of life on the Earth is artificial.
- 23) As appears from the paper, any abnormalities in the structure of the gene and its LCLs (each of the independent informational layers) can result in the organism defectiveness, extinction of species and unpredictable results for the species and Nature at large.
- 24) The paper brings us closer to gene formation laws understanding and consequently to a sensible approach to building up artificial genetic constructions for the treatment of serious diseases and creating new organisms for the benefit of Man and Nature.

References:

1. M. Eingorin «Coding and management in molecular biology», Nizhni Novgorod, The edition NSMA, ISBN № 5-7032-0390-2, 120 pages, April 2001.
2. M. Eingorin “CODING OF A GENE AND GENE ENGINEERING”, «1-st International Congress BIOTECHNOLOGY – state of the art & prospects of development», CONGRESS PROCEEDINGS, October 14-18 2002, Page 14.
3. M. Eingorin, «CODING of the GENETIC TEXTS», http://www.unn.ru/rus/books/eingorin/r1_e.htm
4. A.A. Baranova, M. Eingorin “THE ANALYSIS OF GENETIC CODE DIALECTS USING CODONOGRAM”, «1-st International Congress BIOTECHNOLOGY – state of the art & prospects of development», CONGRESS PROCEEDINGS, October 14-18 2002, Page 15.

The use of these abstracts without due reference to them is not allowed.