**FACTA UNIVERSITATIS**
Series: **Physics, Chemistry and Technology** Vol. 14, N$^{o}$ 3, Special Issue, 2016, pp. 275 - 298
DOI: 10.2298/FUPCT1603275M

# STANDARD GENETIC CODE: *P*-ADIC MODELLING, NUCLEON BALANCES AND SELFSIMILARITY[†]

*UDC* 575.113 : 577.1 +530.145

**Nataša Ž. Mišić**[*]

Lola Institute, Belgrade, Serbia

**Abstract**. *This paper represents the preliminary results and conclusions on the one of fundamental questions of the genetic code related to the underlying selective mechanisms involved in its origin and evolution, in particular their hypothetical different nature, originally considered in [1,2,3]. A novel approach is introduced, based on known arithmetic regularities inside the genetic code, determined by the nucleon balances of amino acids and their divisibility by the decimal number 37 [4]. As a parameter of the genetic code systematization is introduced an aggregate nucleon number of amino acid and cognate codon, while divisibility test is carried out not only by the number 37, but also by 13.7, the selfsimilarity constant of decimal scaling [5]. Relevant nucleon sums were obtained for the most prominent divisions of the standard genetic code (SGC) according to p-adic model of the vertebrate mitochondrial code (VMC) in [6]. The nucleon number divisibility pattern of 37 and 13.7 for the RNA and DNA codon space, as well as for the amino acid space is also analyzed. The obtained results, particularly a general higher divisibility of the nucleon sums by the numbers 37 and 13.7 in SGC than in VMC, as well as a correspondence between the nucleon number divisibility pattern of both the RNA codon space and the amino acid space of SGC, how separately so conjointly, with the code degeneracy pattern, suggest some conclusions: support the hypothesis [1,2,3,7] that the selective driving forces acting during an emergence (an ancient phase) and an evolution (a modern phase) of the genetic code are different, imply the existence of an environmental-dependent stereochemical mechanism throughout the entire period of the genetic code emergence and support a mineral-mediated origin of the genetic code [7,8].*

**Key words**: *genetic code, origin and evolution, stereochemical mechanism, nucleon balances, selfsimilarity*

Received December 1$^{st}$, 2016; accepted December 25$^{th}$ 2016

[†] **Acknowledgement:** The author acknowledges support from the Ministry of Education, Science and Technological Development of Serbia (Projects: TR 45010 and TR 32051). The author would like to thank Professor Branko Dragovich for some discussions.

[*] **E-mail:** nmisic@rcub.bg.ac.rs

## 1. INTRODUCTION

The main flow of genetic information within the biological systems is based on the three sequence-defined biopolymers, nucleic acids DNA/RNA and polypeptides, through the three information processing systems, replication (DNA→DNA), transcription (DNA→RNA) and translation (RNA→polypeptides). Protein synthesis during the translation process is an essential and central biological process in a living cell and at same time the final and most complex step of the genetic information flow. However, microbial phylogenetic studies have revealed that the *translation* process was highly developed at the root of the universal phylogenetic tree, even in comparison to the simpler process of transcription, while a modern type of genome replication mechanism still did not exist at that level [9,10]. The facts that the translation apparatus was the most mature information processing system at the root of the universal tree and that translation process encodes a cell's genotype/phenotype (nucleic acids/proteins) duality make translation one of the main framework for understanding the origin of life.

The simplest abstract description of the complex translation process can be given by the *genetic code* as a map of the set of 64 *codons* (the nucleotide triplets, $N_1N_2N_3$) onto the set of 20 *amino acids* and translation termination *release factors* (prokaryotic RF1 and RF2 or eukaryotic eRF1) (Fig. 1). This strong surjective property of a genetic code mapping implicates that it is a highly degenerate (redundant) code. The degeneracy pattern has a regular form generally determined by the *four*fold degenerate and *two*fold degenerate codon halves (Fig. 1) [11,12]. The genetic code, its codon-amino acid assignment pattern and in particular its degeneracy pattern, is almost universal across all life forms, and a far more prevalent one is referred to as the standard genetic code (SGC; Fig. 1, right). The nonstandard genetic codes are slight variations of the standard code, presented in both nuclear and mitochondrial codes of a wide range of organisms, but mostly those at the bottom of the universal phylogenetic tree (one of the exceptions is very symmetrical and therefore often analyzed the vertebrate mitochondrial code, VMC; Fig. 1, left) [3] (also see NCBI Taxonomy Database).

Beside this principal assignment (degeneracy) pattern, the genetic code is characterized by some others specified by amino acid and codon physico-chemical properties, in particular hydropathy [11,13], amino acid biosynthetic pathways [14,15], classes of aminoacyl-tRNA synthetases [16], amino acid mistranslation and point mutations [17,18,19,20], amino acid frequency in proteins [21], amino acid molecular weighs, i.e. nucleon numbers [22,4,23,24] and others. As shown in the aforementioned articles and some others, all these patterns are, to the some degree, in correspondence to the principal, what is provided a basis for the three main theories of the nature, origin and evolution of genetic code: the stereochemical theory, the coevolution theory and the adaptive theory (for review, see [25,26]). Despite the facts that the central ideas of these theories have been formulated around the time of the genetic code deciphering and that they remained relevant to these days, the numerous subsequent developments have not provided clear and reliable answers on origin and evolution of the genetic code [26]. One of the possible reasons could be that the representative mechanisms were coordinately acting during the code origin. Namely, the compelling evidences support a viewpoint that the origin of SGC and its evolution through the slight variations are the two distinct phases of code evolution, during which dominated the distinct underlying selective mechanisms/driving forces – the ancient phase

with generally less stable environment and more direct association between RNAs and amino acids, and the modern phase with the opposite related characteristics [1,2,3,27]. Moreover, the origin of genetic code is most likely progressively developed through three stages depending on specific domination of one of the three main selective mechanisms, from the initial stereochemical interactions through the metabolic expansion to the final adaptive (error-minimization) adjustment, either in the antagonistic or complementary scenario (see Fig. 5 and explanation in [2]).

Resolving stereochemically nature of the genetic code is extremely important, not only for the answer to why the actual code exactly as it is, but also because it is the main puzzle clue how biotic entities emanated from abiotic components. Since stereochemical mechanism shaped the code in its very emerging and then more or less cooperatively the code is reshaped by another two (metabolic and adaptive) mechanisms, and, as well, that stereochemical domination probably occurred in a completely different circumstances, the revealing of its relics so far resulted by the week and disperse evidences, leaving the stereochemical theory insufficiently grounded (see [2,26]).

Here is proposed that the arithmetical regularities for the representative sums of the aggregate nucleon numbers of amino acids and cognate codons could be so far unknown relic of stereochemical mechanism, and their determination by the selfsimilarity constants of decimal scaling ($37 = 10 \cdot 3.7$, $13.7 \approx 3.7^2$ and $13.7 \approx 3.7+10$), which are also related to the hexagonal lattice, could pull the stereochemical mechanism much deeper in the physical context and support Knight's scenario of complementary evolutionary forces in [2].

## 2. SOME NOTATIONS AND MATHEMATICAL DESCRIPTIONS OF THE GENETIC CODE

The degeneracy pattern in the form of Rumer's principal division on the *four*fold degenerate and *two*fold degenerate codon halves together with their associated amino acids (the rule[1] 4*/2*; Fig. 1) [11], probably represents the most significant signature of driving forces which shaped the genetic code. This pattern 4*/2* is a mainly result of the two Rumer's considerations: 1) a codon separation into its "root" dinucleotide $N_1N_2$ and its "ending" nucleotide $N_3$, and 2) so-called the canonical order C, G, U, A which follows from the nucleobase composition of the dinucleotides $N_1N_2$ according to transition 4*→2* [11,12], where N is any nucleobase, C, G, U and A are respectively the bases cytosine, guanine, uracil and adenine, while a number denotes base position in a codon. The results can be summarized in the three main Rumer's rules which can be briefly described as: 1) the 4* half is determined only by the roots $N_1N_2$, while the 2* half by additional discrimination of pyrimidines $(Y = C, U)$ from purines $(R = G, A)$ at the ending $N_3$, with minor exceptions (footnote 1), 2) if $\mathbf{N} = [C\ G;\ U\ A]$ is a representative matrix of

[1] A division rule of the genetic code into the fourfold and twofold degeneracy halves would be generally denoted as 4* and 2*, respectively. Applied on (multi)sets, this 4*/2* division would be regard as the partition rule on both codon sets and amino acid (multi)sets, how separately so conjointly. With respect to the 2* codon half, it is completely composed of the twofold degenerated codons in VMC, while in SGC there are the two exceptions where a codon quadruplet is not divided in a usual manner 2/2, but as 2/1/1, yet none of these codon quadruplets code more than two amino acids (Fig. 1). The similar is valid for other nonstandard genetic codes, their degeneracy pattern are either of the VMC-like or SGC-like codes. Due to these reasons, the twofold degeneracy can be considered as principal for the 2* half and thus notation adequate.

the base canonical order, then its tensor square $\mathbf{N}^{(2)}$, read by antidiagonals, gives the root canonical order for transition 4*→2* and 3) 4*↔2* transformation is obtained by double transposition C↔A and G↔U known as Rumer's transformation CGUA↔AUGC [11,12]. As the amino bases are $\mathrm{M} = \mathrm{C}, \mathrm{A}$ and the keto $\mathrm{K} = \mathrm{G}, \mathrm{U}$, follows that 4*↔2* transformation is M/K invariant, what corresponds to rotation of the genetic code table by 180º. Using notation for the strong bases $\mathrm{S} = \mathrm{C}, \mathrm{G}$, the week bases $\mathrm{W} = \mathrm{U}, \mathrm{A}$, and total codon set $\overline{C}$, the two degeneracy codon classes can be given in a compact way $4^*(\overline{C}) = \{\mathrm{SS},\ \mathrm{SU},\ \mathrm{WC}\}$ and $2^*(\overline{C}) = \{\mathrm{WW},\ \mathrm{WG},\ \mathrm{SA}\}$ (Fig. 1), showing the prevailing of strong bases in the 4* codon half and the week in the 2* half. Beside this principal division of the genetic code, there are some others which will be exposured through the *p*-adic model (Sec. 3).

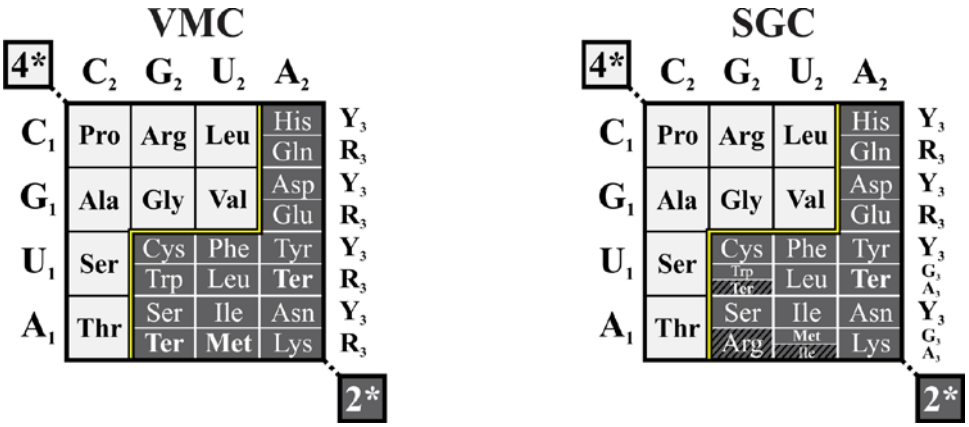


**Fig. 1** Codon/amino acid assignment pattern for the vertebrate mitochondrial code (VMC) and the less symmetrical standard genetic code (SGC) (the hatchings mark the translation alternations between VMC and SGC). Rumer's canonical order of bases is shown for the first and second nucleotide position in a codon, which gives the most compact representation of the two degeneracy classes 4* and 2*, as well as their transformation 4*↔2* under 180º rotation easily obvious.

For the purpose of further presentation, it will be given some mathematical descriptions.

Let denote the set of canonical nucleobases in RNA as $N^{\mathrm{RNA}} = N = \{\mathrm{C}, \mathrm{G}, \mathrm{U}, \mathrm{A}\}$ and the set of nucleobase positions in a codon as $I = \{1, 2, 3\}$. Then the *total codon space* $\overline{C}$ is the cartesian cube of *N*, i.e. $\overline{C} = N^3 = \{n_1 n_2 n_3 : n_i \in N, i \in I\}$, while the (*reduced*) *codon space* $C$ would be $C = \overline{C} \setminus C_{\mathrm{Ter}}$, where $C_{\mathrm{Ter}}$ is the set of termination codons. If $|\cdot|$ denotes cardinality of a set, then $|\overline{C}| = |N^3| = |N|^3 = 4^3$, while $|C|$ depends on the genetic code, e.g. $|C^{\mathrm{SGC}}| = 61$ and $|C^{\mathrm{VMC}}| = 60$. The set of 20 canonical amino acids represents the *amino acid space* $\mathcal{A}$, while its extension by the termination signals is $\overline{\mathcal{A}}$. The *genetic*

*code* is a surjection $g:\overline{C}\to\overline{\mathcal{A}}$, while in narrow sense $g:C\to\mathcal{A}$, what is here more appropriate since the objective is a finding of the intrinsic affinities between codons and amino acids which might have influenced early mutual assignments. Consequently, $\overline{UGN}^{\mathrm{SGC}}=\{\mathrm{UGC,UGG,UGU,UGA}\}\neq UGN^{\mathrm{SGC}}$ since $g(\mathrm{UGA}^{\mathrm{SGC}})=\mathrm{Ter}$ is termination codon, and similar $UGN^{\mathrm{SGC}}\neq UGN^{\mathrm{VMC}}=\{\mathrm{UGC,UGG,UGU,UGA}\}$ since $g(\mathrm{UGA}^{\mathrm{VMC}})=\mathrm{Trp}$ (Fig. 1). Notice different meaning of the notations for a codon set and a codon singlet, e.g. $CGR^{\mathrm{SGC}}=CGR=\{\mathrm{CGG,CGA}\}$, while CGR is a codon CGG or CGA.

Any subset of amino acid space $A_j\subseteq\mathcal{A},\ j\in J=\left\{1,2,3,...,2^{|\mathcal{A}|}\right\}$, has its preimage in a *cognate* subset of codon space $C_j\subseteq C$ such that $g^{-1}\left(A_j\right)=C_j$. Since $g$ is a surjection, it is useful to introduce a multimap $g^{\times}$ which takes into account the repeated elements of $A_j$ so that a multiplicity of amino acid equals to the order of degeneracy of its cognate codons, i.e. $g^{\times}:C_j\to A_j^{\times},\ A_j^{\times}\subseteq\mathcal{A}^{\times}$, where $A_j$ is an *underlying set* of a multiset $A_j^{\times}$ and hence $A_j\neq A_j^{\times}$, as well as $\left|A_j^{\times}\right|=\left|C_j\right|$. For example, $UGN^{\mathrm{SGC}}=\{\mathrm{UGC,UGG,UGU}\}$ has the cognate amino acid sets $g(UGN^{\mathrm{SGC}})=\{\mathrm{Cys,Trp}\}$ and $g^{\times}(UGN^{\mathrm{SGC}})=\{\mathrm{Cys,Cys,Trp}\}$.

A special (multi)set notation for the 4*/2* genetic code division would be given by 4* and 2* as the partition (multi)set rule, i.e. $4^*(C)$ and $2^*(C)$ are respectively the fourfold degenerate and the twofold degenerate codon set, $4^*(\mathcal{A})/2^*(\mathcal{A})$ (or $4^*(\mathcal{A}^{\times})/2^*(\mathcal{A}^{\times})$) are the cognate amino acid (multi)sets. Applying this partition rule on some subset results in the (multi)set intersection of the subset and the corresponding partition of its total set, e.g. $4^*(g^{\times}(YRN))=g^{\times}(YRN)\bigcap 4^*(\mathcal{A}^{\times})=\{\mathrm{Arg,Arg,Arg,Arg}\}$ or $2^*(SSN)=SSN\bigcap 2^*(C)=\varnothing$ since there are no twofold degenerate codons in *SSN*. Due to $4^*(\mathcal{A})\bigcap 2^*(\mathcal{A})=\{\mathrm{Arg,Leu,Ser}\}$, the last notation refers to the multiset $\mathcal{A}^*=4^*(\mathcal{A})\bigcup_+ 2^*(\mathcal{A})$, where $\bigcup_+$ is a multiset sum (an additive union) [28]. Notice that $\left|\mathcal{A}^*\right|=\left|4^*(\mathcal{A})\right|+\left|2^*(\mathcal{A})\right|=8+15=23$ as opposed to $\left|\mathcal{A}\right|=20$.

## 3. *p*-ADIC MODEL OF THE GENETIC CODE AND ITS EUCLIDEAN REPRESENTATION

Biological organisms are based on the information processing systems with a complex, discrete and hierarchical organization. An appropriate theoretical concept and mathematical method for a classification and analysis of bioinformation systems is an ultrametrics [29], since enables not only a description of informational content, but also of informational order and similarity (e.g. a cognate relationship and a contextual closeness). An ultrametric distance is a main tool for such description, and it is defined as a distance which satisfies the strong triangle inequality $d(x,y)\leq\max\{d(x,z),d(z,y)\}$, while a metric space endowed with such distance as an ultrametric space.

Ultrametrics with *p*-adic distances belong to the most elaborated and informative ultrametric spaces, while a *p*-adic modelling of the genetic code and the genome is given

in [6]. Introducing the *p-adic codon space* $C_p$ and the *p-adic amino acid space* $\mathcal{A}_p$ as the subsets of the set $\mathbb{Z}$ of usual integer numbers, the measure of codon-codon similarity and of codon-amino acid assignment closeness were expressed as a distance between the corresponding *p*-adic integers, showing that degeneration of VMC has *p*-adic structure (since all other codes slightly vary from VMC, the result could be generalized) [6].

***p*-Adic distance.** Recall that by Ostrowski's theorem, every nontrivial absolute value on the rational numbers is equivalent to either the usual real absolute value or the *p*-adic absolute value [30]. For a given prime number *p*, $x = \Sigma_{j\geq 0} x_j p^j$ is a $p$-adic integer. Then the *p*-adic absolute value (*p*-adic norm) of a non-zero integer *x* is $|x|_p = p^{-\upsilon(x)}$, where $\upsilon(x)$, so called *p*-adic order, is the highest exponent that $p^{\upsilon(x)}$ divides *x*, while $|0|_p = 0$. Since $\upsilon(x) \in \mathbb{N}_0$ for any integer *x*, follows $|x|_p \leq 1$. For $x = (x_0, x_1, \ldots)$, $y = (y_0, y_1, \ldots) \in \mathbb{Z}$, the *p*-adic distance is defined as

$$d_p(x, y) = \sup_{j\geq 0} \frac{1 - \delta_{x_j y_j}}{p^j} = \frac{1}{p^{\upsilon(x-y)}}, \tag{1}$$

where $\delta_{xy}$ is the Kronecker symbol [31]. From (1) follows that a *p*-adic distance is related to divisibility of $x - y$ by prime *p* (more divisible – smaller distance), what consequently leads to the natural property that two information are closer, i.e. with smaller distance, if they have more equal first digits in their *p*-adic expansion, as well as that digits which come later in the expansion have smaller importance [6]. In the sequel an information space with *p*-adic distance will be called *p*-adic information space $\mathcal{I}_p$.

***p*-Adic codon space $C_p$.** There are two main steps in *p*-adic modelling of codon space: 1) the choosing of an appropriate prime number *p* which will be used as a base for expansion of integers and 2) the identification of an appropriate assignment of *p*-adic digits to the four nucleobases. As the smallest prime number that contains four digits different from zero is $p = 5$ (the use of the digit 0 leads to a nonunique codon representation), the 5-adic integer numbers $c = c_0 c_1 c_2 \equiv c_0 + c_1 5 + c_2 5^2$ for a nonzero digit set $\{1, 2, 3, 4\}$ represent an adequate 5-adic codon space [6], i.e.

$$C_5[4^3] = \left\{ c_0 + c_1 5 + c_2 5^2 : c_i \in 1, 2, 3, 4 \right\} \subset \mathcal{I}_5[5^3]. \tag{2}$$

The corresponding 5-adic distance of any pair of numbers $c, c' \in C_5$ can be:

$$d_5(c, c') = \left| c_0 c_1 c_2 - c_0' c_1' c_2' \right|_5 = \begin{cases} 1, & c_0 \neq c_0' \\ 1/5, & c_0 = c_0', c_1 \neq c_1' \\ 1/25, & c_0 = c_0', c_1 = c_1', c_2 \neq c_2'. \end{cases} \tag{3}$$

Since *p*-adic approach enables the consideration of different distances within the same set of integers, it is possible in the case of $C_5[4^3]$ to quantify the most prominent physico-chemical nucleobase distinction – the pyrimidine and purine type, by introducing 2-adic

distance. There are eight possible connections between nucleobases $\{C, G, U, A\}$ and digits $\{1, 2, 3, 4\}$ which satisfy that 2-adic distance inside the same type is smaller than between different types, i.e. $d_2(C,U) = d_2(G,A) = 1/2$ (Table 4 in [32]). Among these possibilities, there are two which correspond to Rumer's canonical nucleobase order: C,G,U,A and A,U,G,C, and here will be chosen the second one, i.e. $\{A,U,G,C\} = \{1, 2, 3, 4\}$, for the purpose of 2-adic codon-amino acid assignment closeness (originally $\{C,A,U,G\} = \{1, 2, 3, 4\}$ [6,32]).

For all these eight nucleobase-digit assignments, Rumer's general rule for 4*/2* division of codon space can be reformulated that the fourfold degenerate codons are determined by the smallest 5-adic distance (1/25), while the twofold degenerate codons by both 1/25 of 5-adic distance and 1/2 of 2-adic distance [6].

***p*-Adic amino acid space $\mathcal{A}_p$.** A nontrivial *p*-adic representation of a codon-amino acid assignment closeness for $\mathcal{A}_5[20] \subset \mathcal{I}_5[5^3] \setminus \mathcal{C}_5[4^3]$ is obtained when for any amino acid $a = a_0a_1a_2 \in \mathcal{A}_5$ is valid $a_0 \neq 0$, since otherwise 5-adic distance between an amino acid and its cognate codons is maximal, i.e. equal 1.

In the ideal case for the 5-adic distances, the highest closeness can be attained for 16 amino acids in the form $a = a_0a_10 = a_0a_1 \equiv a_0 + a_15$ and $a_0a_1 = c_0c_1$, where $c_0c_1$ is a root dinucleotide part of their cognate codons, and for the rest 4 amino acids in the form $a' = a'_000 = a'_0$ and $a'_0 = c_0$, where $c_0$ is the first base of their cognate codons.

Actually, this ideal model of codon-amino acid assignment is realized in the genetic code with only one exception – Lys in the original assignment [6] or Met in here presented assignment (Table 1) (more detailed in the sequel). Table 1 is principally different from the original one (Table 8 in [6]) in an assignment of the pyrimidine type of bases which enables generally $d_2(a_0a_10, c_0c_12) = d_2(a_0a_10, c_0c_14) = 1/2$ for $a_0a_1 = c_0c_1$, and what is more conserved assignment (there is no anticodon which can make distinction between the codons NNU and NNC, i.e. the codons *c* with $c_3 = 2, 4$) [33]. In [6] is also shown the correspondence of this assignment pattern with a temporal appearance of the canonical amino acids based on [34], as well as 2-adic closeness for the reassignment codons of Leu and Ser, so that generally *p*-adically close codons correspond to the same amino acid.

**Table 1** *p*-Adic representation of canonical amino acid set (according Table 8 in [6]).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 140 **Thr** | 130 Met | 240 ***Ser*** | 230 Cys | 340 **Ala** | 330 **Gly** | 440 **Pro** | 430 ***Arg*** |
| 120 Ile | 110 Asn | 220 Phe | 210 Tyr | 320 **Val** | 310 Asp | 420 ***Leu*** | 410 His |
| 100 Lys | | 200 Trp | | 300 Glu | | 400 Gln | |

– an amino acid *a* which the cognate codons $c \in 4^*(C)$.

– an amino acid *a* which the cognate codons $c \in 2^*(C)$ and $d_5(a,c)=1/25$ (except Lys).

– an amino acid *a'* which the cognate codons $c \in 2^*(C)$ and $d_5(a',c)=1/5$.

***Euclidean representation of p-adic genetic code model.*** A visualizing of the *p*-adic metric is always in the form of a *selfsimilar* structures as are a tree, dendrogram or a fractal, due to the power-law distribution of *p*-adic distances and thus their scale invariance. Here

for the visual representation of *p*-adic information space $\mathcal{I}_5[5^3]$ and its subspaces $\mathcal{C}_5[4^3]$ and $\mathcal{A}_5[20]$, it is used a fractal approach based on a *p*-adic distance representation by usual Euclidean distance, as it is defined in [31]. A formalism will be given for $\mathbb{Z}_p$ and then for $\mathcal{I}_5[5^3] = \mathbb{Z}/5^3\mathbb{Z}$.

Let $V, F \subset \mathbb{R}^n$ and a digit set $P = \{0,1,2,...,p-1\}$, then select an injection $\nu(P) = V$ and define the vector mappings

$$\varphi = \varphi_{\nu,d} : \mathbb{Z}_p \to F, \quad \sum_{j\geq 0} x_j p^j \to \vartheta \sum_{j\geq 0} \frac{\nu(x_j)}{d^{j+1}}, \tag{4}$$

where $\nu(x_j)$ is a digit vector and *d* is a usual Euclidean distance which represents *p*-adic distance $p^{-1}$ and $\vartheta$ is a scaling factor. Since $\mathbb{Z}_p = \bigcup_{x_0 \in P}(x_0 + p\mathbb{Z}_p)$, follows

$$\varphi(\mathbb{Z}_p) = \bigcup_{\nu \in V} \left( \vartheta \frac{\nu}{d} + \frac{1}{d} \varphi(\mathbb{Z}_p) \right), \tag{5}$$

and thus for large enough values of *d*, the image $F = \varphi(\mathbb{Z}_p)$ will be a *disjoint* union of selfsimilar images – a fractal *F* [31].

For a planar representation of 5-adic information space $\mathcal{I}_5[5^3]$ will be $V, F \subset \mathbb{R}^2$, while choosing the digit vectors as $\nu(0) = (0,0)$, $\nu(1) = (-1,1)$, $\nu(2) = (-1,-1)$, $\nu(3) = (1,1)$ and $\nu(4) = (-1,1)$, then $d = 3$ and $\vartheta = 1$, the image $\varphi(\mathcal{I}_5[5^3])$ results in selfsimilar, Cantorian-like set on Fig. 2. The 5-adic genetic model is represented by $\mathcal{G}_5[84] = \mathcal{C}_5[4^3] \cup \mathcal{A}_5[20]$, while $\mathcal{I}_5[125] \setminus \mathcal{G}_5[84]$ is an unused part of information space (light gray numbers on Fig. 2).

If $B_{<r}(x)$ denotes the ball defined by $d_p(x,y) = |x-y|_p < r$ in $\mathbb{Z}_p$, then $\mathcal{I}_5[5^3] = B_{\leq 1}(x)$.

All characteristic properties of finite ultrametric spaces is visible on Fig. 2, such as: 1) any point of a ball is a possible center of the ball $y \in B_{\leq r}(x) \Rightarrow B_{\leq r}(y) = B_{\leq r}(x)$; 2) if two balls have a common point, then one is contained in the other; 3) the set of nonzero distances is always discrete in $\mathbb{R}_{>0}$. Also within *p*-adic metric space, multiplication by *p* in $\mathbb{Z}_p$ is a contracting map

$$d_p(px, py) = \frac{1}{p} d_p(x, y), \tag{6}$$

and hence is continuous. Consequently, for each $y \in 5\mathcal{I}_5[5^2] \subset I_5[5^3]$ follows $y \in B_{\leq 1/5}(x)$, presented by the central ball $B_{\leq 1/5}(x)$ on Fig. 2 (light grey 5-adic numbers). Because of its maximal distance from 5-adic amino acid set, it was excluded from the 5-adic genetic code space, i.e. $\mathcal{G}_5[84] \subset \mathcal{I}_5[125] \setminus 5\mathcal{I}_5[25]$.

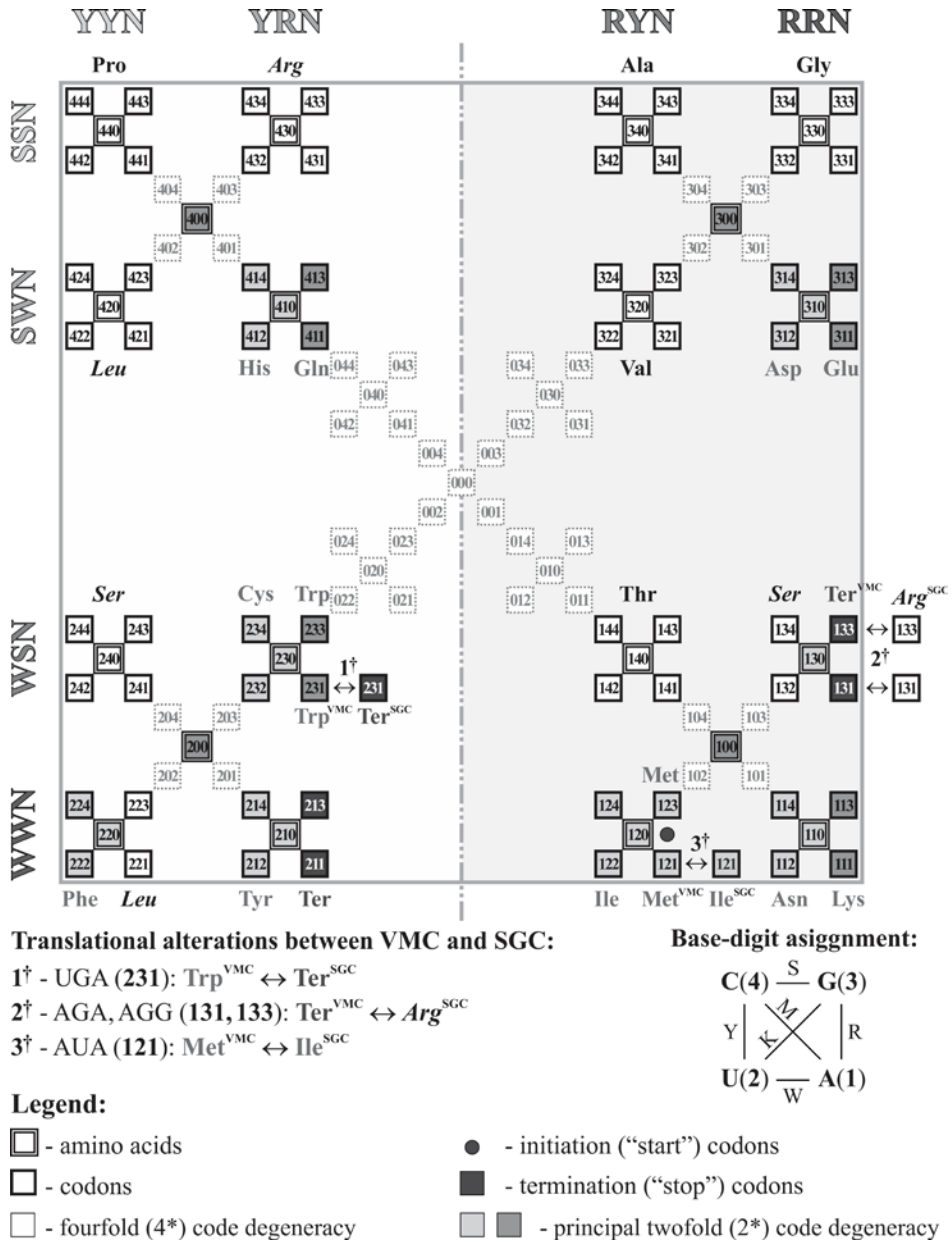


**Fig. 2** Euclidean distance representation of the *p*-adic model of VMC and SGC given by the translational alterations (a colour assignment is the same as on Table 1). The details are described in the text.

The codon arrangement based on Euclidean representation (Fig. 2) can be readily obtained from the mentioned Rumer's nucleobase matrix $\mathbf{N} = [\mathrm{C\ G;\ U\ A}]$ as its tensor cube

$\mathbf{N}^{(3)}$. But *p*-adic code model enables a description of codon clasterization in 4*/2* pattern and a codon-amino acid assignment closeness. Namely, inside the balls $B_{\leq 1/5}(x)$ determined by **4**$c_1c_2$ (***C**NN*) and **3**$c_1c_2$ (***G**NN*) (dominantly 4* class), the four amino acids and their cognate codons are inside the smallest balls $B_{\leq 1/25}(x)$ (the cases $a_0a_1 = c_0c_1$) and the one amino acids with its cognate codons is inside $B_{\leq 1/5}(x)$ (the case $a_0 = c_0$, $a_1 \neq c_1$, Gln for *CNN* and Glu for *GNN*; eq. (3)). Thanks to the reassignment principle and stop codons, the same pattern can be used for a more complex situation inside the balls $B_{\leq 1/5}(x)$ determined by **2**$c_1c_2$ (***U**NN*) and **1**$c_1c_2$ (***A**NN*) (dominantly 2* class). That is fulfilled for **2**$c_1c_2$, but not completely for **1**$c_1c_2$ because of Lys – it would be ideal if Ser and Lys were replaced. Interestingly, such replaced situation destroys the perfect nucleon balance for the aggregate nucleon numbers of $\overline{SSN} = SSN$ and $\overline{WWN}$ (Table 4).

The 2-adic distance shows additional closeness between amino acids and NNY codons since $d_2(c_0c_1 2 - a_0a_1 0) = d_2(c_0c_1 4 - a_0a_1 0) = 1/2$ for $a_0a_1 = c_0c_1$, what is in accordance with the better preservation of assignments between amino acids and NNY codon than NNR. Also $d_2(c_0c_1 3 - a_0 00) = d_2(c_0c_1 1 - a_0 00) = 1/2$ for $a_0 = c_0$ and $c_1 = 1,3$, what is fulfilled for all amino acids coded by NRR if Lys = 100 and Met = 130, what is possible only if the coding as it is or rearranged as in *UNN*. Therefore, for the assignments as in Table 1, the 2-adic distance between amino acids and their closer cognate codons are in all cases 1/2 (for fourfold and threefold degenerate codons, it can be fulfilled only for NNY codons).

The most prominent and regular divisions for Euclidean representation of *p*-adic model of genetic code are emphasized (Fig. 2), such as the divisions according to the first base (*CNN*, *GNN*, *UNN*, *ANN*) and to Y/R, S/W and 4*/2* properties, and which will be used for a determination of the relevant nucleon sums in the next Section (Fig. 3).

## 4. NUCLEON BALANCES

Soon after the genetic code deciphering, an inverse correlation between the size of an amino acid and the number of cognate codons was recognized and then confirmed by introducing an integer-valued parameter – a nucleon number [22]. Further analysis of a nucleon number distribution inside SGC revealed a significant number of arithmetical regularities determined by the decimal number 37 [4,24], as well as some other regularities [36,23,37,38]. Some interesting properties of the number 37 was also shown [4,24,37,38].

A more detailed analysis of a mathematical properties of the number 37, carried out in [37,38,5], indicate an appropriateness of a divisibility testing for the nucleon numbers of the genetic code constituents not only by the number 37, but also by its related number 13.7, which is the selfsimilarity constant of decimal scaling – shortly and roughly: $3.7^2 \approx 13.7$ and $13.7 - 3.7 = 10$ (an exact explanation in Sec. 5).

For the purpose of mathematical description, let introduce notation $\mathrm{n}_X, \mathrm{n}^Y \in \mathbb{N}_0$ for the nucleon number of *i*-tuple nucleobase set, $X \subseteq N^i, i \in I$, and the nucleon number of the amino acid set $Y \subseteq \mathcal{A}^\times$, as well as $\mathrm{n}_X^Y \in \mathbb{N}$ for the aggregate nucleon number of codon set $X \subseteq C$ and its cognate amino acid set $Y = g(X)$ or $g^\times(X)$. The set of

canonical nucleobases in DNA will be emphasized, $N^{\mathrm{DNA}} = \{\mathrm{C},\mathrm{G},\mathrm{T},\mathrm{A}\}$, where T is the thymine. All elements of $N^{\mathrm{RNA}} = N$ or $N^{\mathrm{DNA}}$ would be represented by a nucleobase residue, a nucleobase reduced by one hydrogen atom which is lost during the formation of N-glycosidic bond between a base and a sugar cycle. Consequently, as a representative nucleon number of nucleobase will be considered a nucleon number of free molecules decreased by 1.

Among all arithmetical regularities inside the genetic code, certainly one of the most important is the first discovered regularity of the amino acid nucleon numbers related to Rumer's degeneracy pattern 4*/2* and the 37 divisibility [4]. For the RNA codon space, the 37 divisibility is obtained only for the set of fourfold degenerate codons 4*($C$) [37]. This result is based on previously revealed regularities of DNA and RNA nucleobases

$$\mathrm{n}_{\mathrm{C}} + \mathrm{n}_{\mathrm{G}} = 260 = 7 \cdot 37 + 1 \text{ and } \mathrm{n}_{\mathrm{T}}^{\mathrm{DNA}} + \mathrm{n}_{\mathrm{A}} = 259 = 7 \cdot 37, \tag{7}$$

where $\mathrm{n}_{\mathrm{C}} = 110$, $\mathrm{n}_{\mathrm{G}} = 150$, $\mathrm{n}_{\mathrm{U}}^{\mathrm{RNA}} = \mathrm{n}_{\mathrm{U}} = 111 = 3 \cdot 37$, $\mathrm{n}_{\mathrm{T}}^{\mathrm{DNA}} = 125$ and $\mathrm{n}_{\mathrm{A}} = 134$ [4]. The fact $\mathrm{n}_{4*(C)} = 11988 = 324 \cdot 37$ [37] supports the arithmetical regularities inside the codon space, but inability to find others also denies them. Here it is argued that the main reason for such results basically lies in *a two-level hierarchical structure of the genetic code, and thereby the codon and amino acid space, determined by 4*/2* division with the two different, but related scaling constants for the nucleon sums, 37 and 13.7*.

Comparison of the nucleon regularities according to the 37 and 13.7 divisibility inside RNA and DNA codon space is given with respect to the two criteria of sufficient closeness: 1) a week absolute criterion – a deviation by no more than 2 nucleons, and 2) a strong relative criterion – a deviation less than 3% of value $\mu$ (1 nucleon for 37 and 0.41 for 13.7) (Table 2). Actual values are emphasized (boldfaced) if the closeness is fulfilled at least by the one criterion. Some relevant cases are shown when this closeness is not realized for results interpretation. The same remarks are also valid for Tables 3, 4 and 5.

In sequel it will be listed the results and conclusions based on Table 2.

For nucleobases, the nucleon number divisibility, with respect to both criteria, ranges from according to Rumer's canonical order: C, G, U, (T) and A (actually, the nucleobase A has not a divisibility correspondence). *All these regularities for a divisibility of the nucleobase nucleon numbers in RNA and DNA result in their total nucleon numbers* $\mathrm{n}_N$ *which are closely a product of 37 and 13.7*, i.e. $\mathrm{n}_N^{\mathrm{RNA}} = 505 = 37 \cdot 13.7 - 1.9$ and $\mathrm{n}_N^{\mathrm{DNA}} = 519 = 37 \cdot [13.7] + 1$ (the function $[\cdot]$ rounds to the nerest integer). For nucleobase doublets, the nucleon number divisibility is strongly satisfied for $S = \{\mathrm{C},\mathrm{G}\}$, moderately for $W^{\mathrm{DNA}} = \{\mathrm{T},\mathrm{A}\}$ and weekly for $W^{\mathrm{RNA}} = \{\mathrm{U},\mathrm{A}\}$. It is also interesting that $S$ is better tuned by 13.7 and $W^{\mathrm{DNA}}$ by 37 with difference of 1 nucleon, as well as that their multiplies 7 and 19 are the centered hexagonal numbers $H_2$ and $H_3$, as it is the number 37 ($H_4$), so the total nucleon numbers for canonical DNA base pairing are determined by the first three nontrivial centered hexagonal numbers and 13.7. The set $B^{\mathrm{RNA}} = \{\mathrm{C},\mathrm{G},\mathrm{U}\}$ satisfies almost strong divisibility, what with the regularities for canonical DNA base S/W pairing overall indicates that a replacement of $\mathrm{U}^{\mathrm{RNA}}$ by $\mathrm{T}^{\mathrm{DNA}}$ resulted in a changing of nucleon divisibility distribution – from the hierarchical 4*/2* division in the RNA genetic code to

**Table 2** Comparison of the nucleon distribution for the RNA and DNA nucleobase singlets, doublets and triplets, as well as a deviation of the nearest multiplies of 37 and 13.7 from the actual nucleon values.

| | **SGC (RNA)** | | | | | **SGC (DNA)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mu$ = **37** | | $\mu$ = **13.7** | | | $\mu$ = **37** | | $\mu$ = **13.7** | | |
| | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | Actual | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | Actual |
| $n_C$ | 3 | **111** | 8 | **109.6** | **110** | 3 | **111** | 8 | **109.6** | **110** |
| $n_G$ | 4 | **148** | 11 | **150.7** | **150** | 4 | **148** | 11 | **150.7** | **150** |
| $n_U^{RNA}$, $n_T^{DNA}$ | 3 | **111**† | 8 | **109.6** | **111**† | 3 | 111 (-14) | 9 | **123.3** | **125** |
| $n_A$ | 4 | 148 (+14) | 10 | 137 (+3) | 134 | 4 | 148 (+14) | 10 | 137 (+3) | 134 |
| $n_S$ | 7 | **259** | 19 | **260.3** | **260** | 7 | **259**† | 19 | **260.3** | **260**† |
| $n_W$ | 7 | 259 (+14) | 18 | **246.6** | **245** | 7 | **259**† | 19 | **260.3** | **259**† |
| $n_B$ | 10 | **370** | 27 | **369.9** | **371** | 10 | 370 (-15) | 28 | **383.6** | **385** |
| $n_N$ | 14 | 518 (+13) | 37 | **506.9** | **505** | 14 | **518** | 38 | **520.6** | **519** |
| $n_{NN}$ | 109 | 4033 (-7) | 295 | **4041.5** | **4040** | 112 | 4144 (-8) | 303 | **4151.1** | **4152** |
| $n_{\overline{NNN}} = n_{\overline{C}}$ | 655 | 24235 (-5) | 1769 | 24235.3 (-4.7) | 24240 | 673 | 24901 (-11) | 1818 | 24906.6 (-5.4) | 24912 |
| $n_{NNN} = n_C$ | 624 | 23088 (+17) | 1684 | **23070.8** | **23071** | 641 | 23717 (+16) | 1730 | **23701** | **23701** |
| $n_C / \lvert C \rvert$ | 10 | 370 (-9) | 28 | 383.6 (+4.6) | 379 | 11 | 407 (-18) | 28 | 383.6 (-5.4) | 389 |
| $n_C / \lvert \overline{C} \rvert$ | 10 | 370 (+10) | 26 | 356.2 (-3.8) | 360 | 10 | **370** | 27 | **369.9** | **370** |

† – the results revealed by Shcherbak [4,24].

*B* – the set of not A nucleobases, i.e. the set of C, G and U.

$[x,\mu]$ – a nearest integer function, rounds to the nearest integer multiple of $\mu$.

**boldface** – a calculated value $[x,\mu]\cdot\mu$ which not deviates from the actual one by more than 2 nucleons.

☐ – an actual value or a calculated value $[x,\mu]\cdot\mu$ which deviates from the actual one by less than 3% of a value $\mu$.

the uniform S/W division in the DNA genetic code. It means that *the U↔T replacement is such that, in terms of the nucleon numbers, makes RNA nucleobase space better tuned for a codon degeneracy, while DNA nucleobase space better tuned for a strain pairing*. It also imposed a conclusion that the genetic code (SGC) originated from an RNA world or more precisely, as further results indicated (Table 3), from an affinity between an RNA and amino acid world, in accordance with the stereochemical theory. This U↔T replacement in the RNA and DNA codon table of SGC also results in a changing of 48 positions in 37 codons with a total nucleon difference by $672 = 49 \cdot 13.7 + 0.7$, while in their reduced codon tables by termination codons results in a changing of 45 positions in 34 codons with a total nucleon difference by $630 = 17 \cdot 37 + 1 = 46 \cdot 13.7 + 0.2$. The last results are a consequence of the nucleon number divisibility for both RNA and DNA nucleobase triplets – codons, which shows *no correspondence for the total codon spaces, but only for the reduced codon spaces of SGC by 3 termination codons* when, with respect to the 13.7

divisibility, gives a strong correspondence with the actual nucleon values, $23071 = 1684 \cdot 13.7 + 0.2$ for RNA and exact $23701 = 1730 \cdot 13.7$ for DNA. Restriction of the nucleon regularities just on the reduced RNA and DNA codon tables is again a strong argument for the sterochemical theory which is based on affinity between codons and amino acids, and not between codons and release factors. Moreover, it further indicates that a stereochemical mechanism were in effect until the final genetic code shaping, what supports Knight's scenario of complementary evolutionary forces [2]. This also gave a base for all further analysis to be considered on the reduced codon tables (Tables 3 and 4), what a detailed analysis confirmed as justified (not presented here). Finally, an average nucleon number of a codon, $n_C/|C| = n_C/61$, is 379 for RNA and 389 for DNA, which is not so far from $370 = 10 \cdot 37 \approx 27 \cdot 13.7$, while in a special case $n_C/64$ is exactly 370 for DNA.

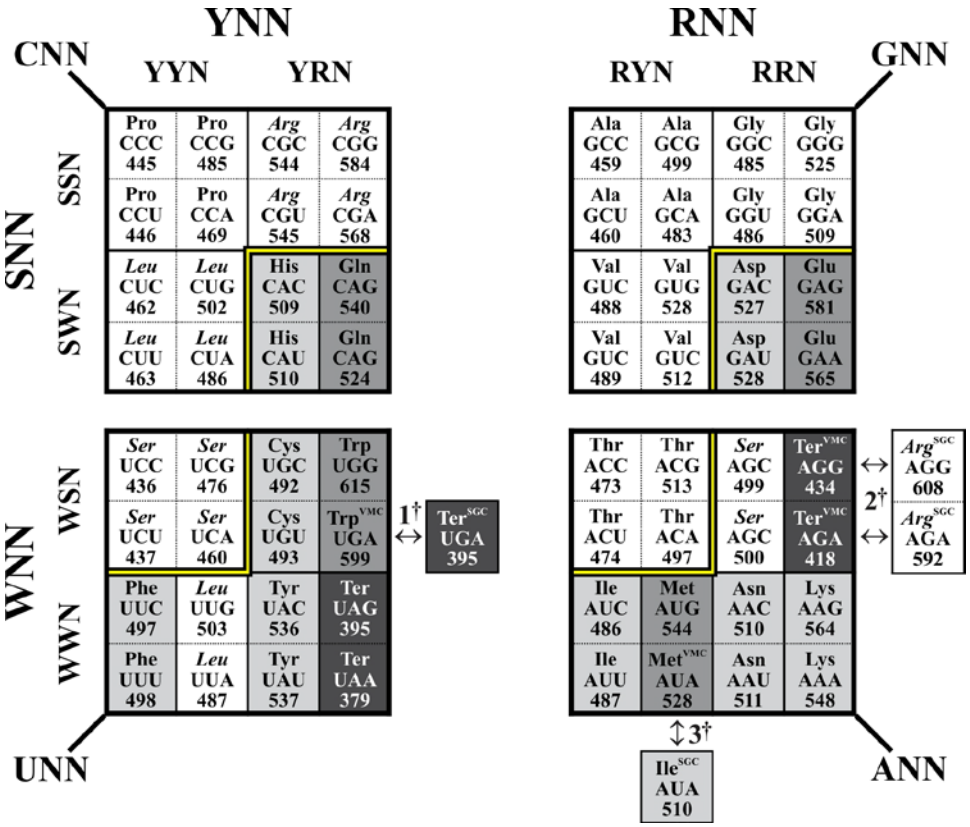


**Fig. 3** A distribution of the aggregate nucleon numbers for Euclidean representation of the *p*-adic model of VMC and SGC given by the translational alterations as on Fig. 2. The most regular division for such genetic code representation are emphasized.

A distribution of the aggregate nucleon numbers of a codon and its cognate amino acid, $n_c^a$, for Euclidean representation of the *p*-adic model of VMC and SGC is shown on Fig. 3. The nucleon numbers only of (*free*) amino acids is given in [4]. The most prominent and

regular divisions for this genetic code representation are emphasized, such as the divisions according to the first base (*CNN*, *GNN*, *UNN*, *ANN*) and to Y/R, S/W and 4*/2* properties.

Comparison of the nucleon divisibility distribution for the codon and amino acid sets in VMC and SGC according to 4*/2* degeneracy pattern is presented on Table 3. It is easily evident a higher nucleon divisibility by 37 and 13.7 in SCG then in VMC, and what is more important – all nucleon numbers which correspond with the 37 and 13.7 multiplies according to a strong and weak criterion are those originating from SGC, except in the case of an average aggregate nucleon number for the reduced genetic code tables (the last raw in Table 3), which values in both cases are very close to the product of scaling constants 37 and 13.7, i.e. $37 \cdot 13.7 = 506.9$ (similarly to $n_N$, Table 2). In another words, all changes in VMC have resulted negatively in terms of the nucleon regularities except the average aggregate nucleon number. Accordingly, the results will be interpreted only for SGC.

**Table 3** Comparison of the nucleon distribution for the codon and amino acid sets in VMC and SGC according to 4*/2* degeneracy pattern, as well as a deviation of the nearest multiplies of 37 and 13.7 from the actual nucleon values.

| | **VMC** | | | | | **SGC** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mu$ = **37** | | $\mu$ = **13.7** | | | $\mu$ = **37** | | $\mu$ = **13.7** | | |
| | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | Actual | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | Actual |
| $n^{4*(\mathcal{A})}$ | 25 | **925** | 68 | 931.6 (+6.6) | **925** | 25 | **925**† | 68 | 931.6 (+6.6) | **925**† |
| $n^{2*(\mathcal{A})}$ | 60 | **2220** | 162 | **2219.4** | **2220** | 60 | **2220**† | 162 | **2219.4** | **2220**† |
| $n^{\mathcal{A}*}$ | 85 | **3145** | 230 | 3151 (+6) | **3145** | 85 | **3145**† | 230 | 3151 (+6) | **3145**† |
| $n^{\mathcal{A}*}/\lvert\mathcal{A}*\rvert$ | 4 | 148 (+11) | 10 | **137** | **[136.74]** | 4 | 148 (+11) | 10 | **137** | **[136.74]** |
| $n^{4*(\mathcal{A}^c)}$ | 100 | **3700** | 270 | **3699** | **3700** | 100 | **3700**† | 270 | **3699** | **3700**† |
| $n^{2*(\mathcal{A}^c)}$ | 111 | 4107 (+15) | 299 | 4096.3 (-4.3) | 4092 | 114 | **4218**† | 308 | **4219.6** | **4218**† |
| $\Delta$ | 11 | 407 (+15) | 29 | 397.3 (-4.3) | 392 | 14 | **518** | 38 | ***520,6 (+2.6)*** | **518** |
| $n^{\mathcal{A}^c}$ | 211 | 7807 (+15) | 569 | 7795.3 (-3.3) | 7792 | 214 | **7918**† | 578 | **7918.6** | **7918**† |
| $n_{4*(C)}$ | 324 | **11988** | 875 | **11987.5** | **11988** | 324 | **11988** | 875 | **11987.5** | **11988** |
| $n_{2*(C)}$ | 287 | 10619 (-7) | 776 | 10631.2 (+5.2) | 10626 | 300 | 11100 (+17) | 809 | **11083.3** | **11083** |
| $n_C$ | 611 | 22607 (-7) | 1651 | 22618.7 (+4,7) | 22614 | 624 | 23088 (+17) | 1684 | **23070.8** | **23071** |
| $n_C^{\mathcal{A}^c}$ | 822 | 30414 (+8) | 2219 | 30400.3 (-5.7) | 30406 | 838 | 31006 (+17) | 2262 | **30989.4** | **30989** |
| $n_C^{\mathcal{A}^c}/\lvert C\rvert$ | 14 | 518 (+11) | 37 | **506.9** | **[506.77]** | 14 | 518 (+10) | 37 | **506.9** | **[508.01]** |

† – the results revealed by Shcherbak [4,24].

$\Delta$ – a difference of the first two above values.

$[x,\mu]$ – a nearest integer function, rounds to the nearest integer multiple of $\mu$.

**boldface** – a calculated value $[x,\mu]\cdot\mu$ which not deviates from the actual one by more than 2 nucleons.

– an actual value or a calculated value $[x,\mu]\cdot\mu$ which deviates from the actual one by less than 3% of a value $\mu$.

– a sum of the appropriate values inside a block or a difference $\Delta$.

A difference between the multiset $\mathcal{A}^*$ and the set $\mathcal{A}$ is in repeating of 3 amino acids, what results in $n^{Ser,Leu,Arg} = 105+131+174 = 410 = 30\cdot 13.7 - 1$ ($n^{\mathcal{A}^*} - n^{\mathcal{A}} = 3145 - 2735 = 410$), giving also that an average nucleon number of these 3 amino acids is closely 137, similarly as for $\mathcal{A}$ and thus $\mathcal{A}^*$ (Table 3). Taking into account the amino acid ionized/protonated forms gives an exact relation $n^{Ser,Leu,Arg^{(+1)}} = 105+131+175 = 411 = 30\cdot 13.7$ and an exact value 137 for their average nucleon number, while for $\mathcal{A}$ and $\mathcal{A}^*$ will be $3146 - 2735 = 411$.

Compared with all other considered divisions (Fig. 3), the 4*/2* division reflects the most faithful image of 37 and 13.7 divisibility, not only because almost strict correspondence, but also because *all regularities were coherently realized, how separately at the level of codons and amino acids, so consequently together*. Moreover, the nucleon regularities of amino acids were realized both for the multisets $\mathcal{A}^*$ and $\mathcal{A}^{\times}$ [4,24], but when both criteria are considered then 4*/2* division in the set $\mathcal{A}^{\times}$ shows a better agreement. Namely, the nucleon numbers of both $4^*(\mathcal{A}^{\times})$ and $2^*(\mathcal{A}^{\times})$ are exactly determined by 37 and in a lesser degree by 13.7, while their nucleon difference shows the same principle as for $n_N$ (Table 2) and $n_C^{\mathcal{A}^{\times}}/|C|$ (Table 3), a closeness with $37\cdot[13.7] = 518$ or $37\cdot 13.7 = 506.9$. A compelling evidences of the decimal scaling by 37 (3.7) in SGC, eq. (14), is that its most representative set $4^*(\mathcal{A}^{\times})$ has exactly 3700 nucleons, as well as that a distribution of nucleons for $\mathcal{A}^{\times}$, $4^*(\mathcal{A}^{\times})$ and $2^*(\mathcal{A}^{\times})$ fulfils (22) which directly follows from the generating equation (13) of the number 37.

In the reduced codon space, the difference in the fine tuning of nucleon numbers for the sets $4^*(C)$ and $2^*(C)$ is obvious, the first is tuned by 37 and the second by 13.7, what overall results in the same global tuning principle of SGC, $n_{4^*(C)}^{4^*(\mathcal{A}^{\times})} = 15688 = 424\cdot 37$ and $n_{2^*(C)}^{2^*(\mathcal{A}^{\times})} = 15301 = 1117\cdot 13.7 - 1.9$ (Table 3). As the most reasonable explanation is imposed the previous statement about possibility that SGC, and consequently its modified versions, is organized as *a nested hierarchical structure determined by its degeneracy pattern (the 4*/2* division), what on the representative nucleon numbers of its constituents is reflected as their clustering for the scaled values of the same number 3.7, i.e. 3.7·10 and $3.7^2$*. The best support for this hypothesis is the very fact that the number 3.7 is derived from a decimal scaling (12)-(14).

Almost all regularities (Table 4) have a strong correspondence by one of the scaling values, and that most for 13.7. In particular, the almost exact divisibility by 13.7 is attained for *YNN/RNN* division of SGC, what has special importance in the context of the fact that this division makes exact half of 4*/2* (Fig. 3), which particular sets $4^*(YNN)$, $4^*(RNN)$, $2^*(YNN)$ and $2^*(RNN)$ have a divisibility correspondence of the aggregate numbers for given two criteria in the order 1→4 (Table 5). Overall, the sets $4^*(YNN)$ and $2^*(YNN)$ are better tuned by 37, while $4^*(RNN)$ and $2^*(RNN)$ by 13.7. In respect to $4^*(\mathcal{A}^{\times})$, the *YNN/RNN* division results in the nucleon number division as $3700 = 2100 + 1600$. In the context of codon family (codon quadruplets), generally the best tuned are CCN and CGN which code Pro and Ala, respectively (Fig. 3). Their special place in *p*-adic model of SGC has also special place in a nucleon divisibility, i.e. $n^{\{Pro^{\times},Ala^{\times}\}} = 4(115+89) = 816 = 22\cdot 37 + 2$. Interestingly, those two amino acid are sterochemically the most untypical – Pro, and the

most typical – Ala. Namely, Pro is the only imino acid, which under biological condition has the protonated form and thus actually is an amino acid, as well as the only *cyclic* amino acid which consequently induce an exceptional conformational rigidity compared to other amino acids. Contrary, Ala, with its methyl group $-CH_3$ in side-chain, is stereochemically representative amino acid of the 16-membered amino acid group with a methylene bridge $-CH_2-$ in side-chain. The fact that such sterochemically special amino acids has a special place in the *p*-dic SGC (Fig. 3) and the nucleon regularities, gives once again a support to the sterochemical theory. Finally, the difference of the aggregate nucleon numbers for the *RNN/YNN* division has the value $[3.7^3] = [50.653] = 51$, which means that the number 3.7, eq. (14), not only appears as a scaling factor of the first and second order (3.7 and $3.7^2$), but also of the third order.

**Table 4** The nucleon distribution of SGC for the divisions according to its *p*-adic model (Fig. 3), as well as a deviation of the nearest multiplies of 37 and 13.7 from the actual nucleon values (the notations are the same as for Tables 2 and 3).

| | **SGC** | | | | |
|---|---|---|---|---|---|
| | $\mu$ = **37** | | $\mu$ = **13.7** | | |
| | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | $[x,\mu]$ | $[x,\mu]\cdot\mu$ | Actual |
| $n_{CNN}^{g^\times(CNN)}$ | 218 | 8066 (-16) | 590 | **8083** | **8082** |
| $n_{GNN}^{g^\times(GNN)}$ | 220 | 8140 (+16) | 593 | **8124.1** | **8124** |
| $n_{UNN}^{g^\times(UNN)}$ | 175 | 6475 (+8) | 472 | **6466.4** | **6467** |
| $n_{ANN}^{g^\times(ANN)}$ | 225 | 8325 (+9) | 607 | **8315.9** | **8316** |
| $n_{YNN}^{g^\times(YNN)}$ | 393 | 14541 (-8) | 1062 | **14549.4** | **14549** |
| $n_{RNN}^{g^\times(RNN)}$ | 444 | 16428 (-12) | 1200 | **16440** | **16440** |
| $\Delta$ | 51 | 1887 (-4) | 138 | **1890.6** | **1891** |
| $n_{SSN}^{g^\times(SSN)}$ | 216 | **7992** | 583 | 7987.1 (-4.9) | **7992** |
| $n_{SWN}^{g^\times(SWN)}$ | 222 | **8214** | 600 | 8220 (+6) | **8214** |
| $n_{WSN}^{g^\times(WSN)}$ | 204 | 7548 (-17) | 552 | ***7562.4 (-2.6)*** | 7565 |
| $n_{WWN}^{g^\times(WWN)}$ | 195 | 7215 (-3) | 527 | **7219.9** | **7218** |
| $n_{\overline{WSN}}^{g^\times(WSN)}$ | 215 | 7955 (-5) | 581 | **7959.7** | **7960** |
| $n_{\overline{WWN}}^{g^\times(WWN)}$ | 216 | **7992** | 583 | 7987.1 (-4.9) | **7992** |
| $n_{SNN}^{g^\times(SNN)}$ | 438 | **16206** | 1183 | **16207.1** | **16206** |
| $n_{WNN}^{g^\times(WNN)}$ | 400 | 14800 (+17) | 1079 | **14782.3** | **14783** |
| $\Delta$ | 38 | 1406 (-17) | 104 | **1424.8** | **1423** |

From the aspect of S/W properties (Table 4), the aggregate numbers of the sets *SSN* and *SWN* are the exact 37 multiplies, and hence their sum, what correlates with the fact that 6 of total 8 codon quadruplets of *SNN* belong to the 4* half which is exactly determined also by 37 multiply. A complementary consideration can be given for the sets *WSN* and *WWN*, as well as for their total set *WNN*, which are better tuned by 13.7 and whose 6 of total 8 codon quadruplets of *WNN* belong to the 2* half which is closely determined also by 13.7 multiply. Peculiarity of *WSN* and *WWN* are better tuning of the aggregate nucleon numbers with a counting of termination codons $n_{\overline{WSN}}^{g^{\times}(WSN)} = 581 \cdot 37 = 7959.7 + 0.3 = 7960$ and $n_{\overline{WWN}}^{g^{\times}(WWN)} = 216 \cdot 37 = 7992 = n_{\overline{SSN}}^{g^{\times}(SSN)} = n_{SSN}^{g^{\times}(SSN)}$. The last result is interesting since *SSN* is all 4* codon set and hence the simplest of all prominent sets, while $\overline{WWN}$ is all 2* codon set and hence the most complex of all prominent sets (Fig. 3), but yet they have the same aggregate nucleon number.

**Table 5** The nucleon distribution of the *p*-adic SGC for the composition of divisions *YNN/RNN* and 4*/2* (Fig. 3), as well as a deviation of the nearest multiplies of 37 and 13.7 from the actual nucleon values (the notations are the same as for Tables 2, 3 and 4).

| | **SGC** | | | | |
|---|---|---|---|---|---|
| | $\mu = 37$ | | $\mu = 13.7$ | | |
| | $[x, \mu]$ | $[x, \mu]\cdot\mu$ | $[x, \mu]$ | $[x, \mu]\cdot\mu$ | Actual |
| (1) $n_{4*(YNN)}^{g^{\times}(4*(YNN))}$ | 211 | **7807** | 570 | **7809** | **7808** |
| (2) $n_{4*(RNN)}^{g^{\times}(4*(RNN))}$ | 213 | **7881** | 575 | ***7877.5 (-2.5)*** | **7880** |
| (3) $n_{2*(YNN)}^{g^{\times}(2*(YNN))}$ | 182 | 6734 (-7) | 492 | **6740.4** | **6741** |
| (4) $n_{2*(RNN)}^{g^{\times}(2*(RNN))}$ | 231 | 8547 (-13) | 625 | ***8562.5 (+2.5)*** | 8560 |
| $\Delta(1,2)$ | 2 | **74** | 5 | 68.5 (-3.5) | **72** |
| $\Delta(1,3)$ | 29 | 1073 (+6) | 78 | **1068.6** | **1067** |

Codon space is analyzed for the nucleobase residues, while for their free molecules with an 1 nucleon higher nucleon numbers, the whole codon table results in a total increasing of nucleons for $3 \cdot 64 = 14 \cdot 13.7 + 0.2 = 192$, but despite this regularity almost *none* of the presented regularities for the aggregate nucleon numbers are preserved.

All previous considerations concerning the amino acid nucleon numbers are based on their free molecules. Since the nucleon balance regularities are founded [23] also for the amino acid residues and their ionized/protonated forms ($Asp^{(-1)}$, $Glu^{(-1)}$, $Lys^{(+1)}$, $Arg^{(+1)}$; since His is the only amino acid whose side-chain can switch from an unprotonated to a protonated state under neutral pH conditions due to the $pK_a$ value of 6.0 of its side-chain, in the paper [23] is taken a value as for a free molecule), these results will be also analyzed in terms of their divisibility.

Since a forming peptide bond results in release of a molecule of water ($n^{H_2O} = 18$), in an amino acid main-chain remains $74 - 18 = 56 = 4 \cdot 13.7 + 1.2$ nucleons. If ñ is a nucleon

number for the residual molecules then will be $\tilde{n}^{\mathcal{A}} = 2357 = 172 \cdot 13.7 + 0.6$ while its division on main-chain and side-chain sets gives respectively 1119 and 1256 with their exact difference 137 (see Table 1 in [23]). Moreover, similarly to (22), $(137 = 64 + 73) \cdot 17.2$ roughly describe this nucleon distribution. For $\mathcal{A}^{\times}$, the main-chain and side-chain nucleon numbers are exact halves of total set with value $\tilde{n}^{\mathcal{A}^{\times}} / 2 = 3412 = 249 \cdot 13.7 + 0.7$. The inclusion of nonstandard codes in the analysis of divergence between the amino acid main-chain and side-chain nucleon numbers for all codon table shows that SGC is the only genetic code with the null divergence [23]. The same analysis of proteins in the wide range of species showed that a nucleon distribution in coded proteins is correlated with genomic base composition, as well as that on average the total main-chain and side-chain nucleon numbers of proteins have approximately equal values [23]. As one of the possible reasons of such tuning of the nucleon numbers, and thus a mass, in proteins is suggested an *optimization of their dynamical properties* with the final concluding remark that, "in summary, whatever the driving force behind the observed pattern, it seems likely that a genetic code based on hydrophobicity and mass balance holds a central place in the evolution of genome at the chemical level" [23]. This viewpoint is supported here, with an additional considering of nucleon divisibility regularities.

In the end, a simple question – why 13.7, not exactly $3.7^2$=13.69? A simple answer is 13.7 gives a better correspondence with the actual values, as well as very recognizable pattern – where better tunes 37 there worse tunes 13.7 and vice versa. Typical examples are a better tuning of the amino acids and the 4* half by 37, while the codons and the 2* half by 13.7. A more advanced answer may follow from the *unit* difference of their higher-scale values 1369 and 1370 or 3700 and $3699 = 270 \cdot 13.7$ (while $270 \cdot 13.69 = 3696.3 = 3700 - 2.7$), what means that these two constants enable both a scaling for the powers of 10 and an unit shifting. To understand deeper reasons, it is necessary to consider the properties of these two numbers.

## 5. SELFSIMILAR AND SCALING PROPERTIES OF THE NUMBERS 37 AND 13.7

Besides revealing the distribution patterns of nucleon sums divisibility by the numbers 37 and 13.7 and their understanding in the context of symmetry and physico-chemical properties of the genetic code constituents, it is necessary to understand a mathematical properties of the numbers 37 and 13.7 and their potential relation with a physical reality. In sequel, it will be given some mathematical arguments which overall indicate that those numbers are related to the selfsimilar symmetry and the scaling by powers of 10.

A generalization of the number 37 for a different base of numeral system *q* and a digit multiplicity *m* showed that its basic property is an *equidistant* cycling digit property (an equidistance both for the multipliers and digits) and that its generalized numbers, *Shcherbak numbers* ($\mathcal{S}$), have a simple form

$$\mathcal{S}_m(q) = \frac{R_m(q)}{m}, \text{ for } m \mid q - 1, \tag{8}$$

where $R_m(q)=\Sigma_{k=0}^{m-1}q^k=11...11_q$ is the generalized Niven (Harshad) repunit of length *m* and $m,q\in\mathbb{N}_{\geq 2}$ [36]. Since for a positive integer *l* when *l*|*m* then $R_l|R_m$, follows that the *irreducible* $\mathcal{S}$ have the form $\mathcal{S}_p(q)=R_p(q)/p$, where *p* is prime. The fact that a repunit polynomial $R_p(q)$ also represents an irreducible cyclotomic polynomial $\Phi_p(q)$ which roots are all $p^{\text{th}}$ primitive roots of unity and lie on the unit circle in the complex plane [40], gives some relation $\mathcal{S}$ with the regular geometrical patterns (e.g. all $\mathcal{S}_3(q), 3|q-1$, are the centered hexagonal numbers [37,38,5]).

An extension of $\mathcal{S}$ within same numeral system results in the *q*-scaled numbers $\tilde{\mathcal{S}}$,

$$\tilde{\mathcal{S}}_{rp}(q)=\frac{R_{rp}(q)}{pR_r(q)}=\frac{R_p(q^r)}{p}=\mathcal{S}_p(q^r),\quad r\in\mathbb{N}, \tag{9}$$

which have similar numerical, arithmetical and geometrical properties, while in a special case when $r=p$ and $q=p^2+1$, then

$$\mathcal{S}_{q-1}(q)=\mathcal{S}_{p^2}(q)=\mathcal{S}_p(q)\mathcal{S}_p(q^p)=\mathcal{S}_p(q)\tilde{\mathcal{S}}_{p^2}(q) \tag{10}$$

(e.g. for $q=3^2+1=10$ follows $\mathcal{S}_{3^2}(10)=12345679=37\cdot 333667$, were $\tilde{\mathcal{S}}_{3^2}(10)=333667_{10}$ $=\overline{333}\,\overline{667}_{10^3}=\mathcal{S}_3(10^3)$, if we adopt notation for the digits as $1,2,...,9,\overline{10},\overline{11},...\overline{999}$) [38]. A general form of (10) is

$$\mathcal{S}_{p^s}(q)=\mathcal{S}_{p^{s-1}}(q)\mathcal{S}_p(q^{p^{s-1}})=\Pi_{k=1}^{s-1}\mathcal{S}_p(q^{p^k})=\Pi_{k=1}^{s-1}\tilde{\mathcal{S}}_{p^s}(q),\quad s\in\mathbb{N}_{\geq 2}, \tag{11}$$

In a special case when a numeral system has the base $q=p^s+1$, its highest Shcherbak number has the form $\mathcal{S}_{p^s}(q)=123...\overline{q-3}\,\overline{q-1}=\Pi_{k=1}^{s-1}\tilde{\mathcal{S}}_{p^s}(q)$, which is related to both the numeration process, as well as $R_\infty^2(q)=(123...\overline{q-3}\,\overline{q-1})^\infty$, and to the nested product of selfsimilar *q*-scaled numbers $\tilde{\mathcal{S}}_{p^s}(q)=\mathcal{S}_p(q^{p^{s-1}})$, what is the case of decimal system [5,38]. These selfsimilar properties of Shcherbak numbers are once again a consequence of the properties of correspondent cyclotomic polynomials. i.e. $\Phi_{p^s}(q)=\Phi_p(q^{p^{s-1}})$ [38]. Since $q^{km}\equiv 1 \bmod \mathcal{S}_m(q), m|q-1, k\in\mathbb{N}$, e.g.

$$10^{3k}-1=3_k6_{k-1}7\cdot 29_{k-1}7=\tilde{\mathcal{S}}_{3k}(10)\cdot 29_{k-1}7, \tag{12}$$

where the indices denote the number of a *digit* repetition, it is interesting to consider an existence of the similar numbers that give accurate *q*-scaling, in particular for $q=10$.

From the general conditions $\Psi_{(s)}=10\Psi_{(s-1)}$, $\psi_{(s)}=10\psi_{(s-1)}, s\in\mathbb{Z}$, and particular $\Psi_{(1)}\psi_{(1)}=10^3$ and $\Psi_{(1)}-\psi_{(1)}=10$, the numbers $\Psi_{(s)}$ and $\psi_{(s)}$ can be obtain as the positive solutions of the polynomial equations [5]

$$\Psi_{(s)}^2 - 10^s \Psi_{(s)} - 10^{2s+1} = 0\,, \tag{13}$$

$$\psi_{(s)}^2 + 10^s \psi_{(s)} - 10^{2s+1} = 0\,, \tag{14}$$

and the values of these irrationals are

$$\Psi_{(s)} = 3.7015... \cdot 10^s \text{ and } \psi_{(s)} = 2.7015... \cdot 10^s\,. \tag{15}$$

Concretely, (15) gives $\Psi_{(1)}\psi_{(1)} = 37.015... \cdot 27.015... = 10^3$, while eq. (12) $37 \cdot 27 = 10^3 - 1$. If $\Psi_{(0)} = \Psi = 3.7015...$ and $\psi_{(0)} = \psi = 2.7015...$, then

$$\Psi_{(s)}\psi_{(s)} = 10^{2s}\Psi\psi = 10^{2s+1}\,, \tag{16}$$

$$\Psi_{(s)} / \psi_{(s)} = \Psi/\psi = \Psi^2/10 = \Psi/10 + 1 = 1.37015...\,, \tag{17}$$

$$\Psi_{(s)} - \psi_{(s)} = 10^s(\Psi - \psi) = 10^s \text{ and} \tag{18}$$

$$\Psi_{(s)} + \psi_{(s)} = 10^s(\Psi + \psi) = 6.40312... \cdot 10^s\,. \tag{19}$$

Using the functions round nearest $\lfloor \cdot \rceil$ and round up $\lceil \cdot \rceil$, from (19) for $s = 1$ follows $[\Psi_{(1)}] + [\psi_{(1)}] = [10\Psi] + [10\psi] = 27 + 37 = 64$, and further

$$\left\lfloor \mathbf{64} \cdot 10\Psi^2 - 23\Psi \right\rceil = \left\lceil \mathbf{27} \cdot 10\Psi^2 \right\rceil + \left\lfloor \mathbf{37} \cdot 10\Psi^2 - 23\Psi \right\rceil, \tag{20}$$

$$7918 = 3700 + 4218 \quad \Leftrightarrow \quad 214 \cdot \mathbf{37} = 100 \cdot \mathbf{37} + 114 \cdot \mathbf{37}\,, \tag{21}$$

where $\mathrm{n}^{\mathcal{A}^\times} = 7918$, $\mathrm{n}^{4*(\mathcal{A}^\times)} = 3700$ and $\mathrm{n}^{2*(\mathcal{A}^\times)} = 4218$ (Table 3). Better agreement it can be obtain from (13) and (14) as $[10\Psi^2] - [10\psi^2] = 137 - 73 = 64 = [10(\Psi + \psi)]$,

$$\left\lfloor \mathbf{137} \cdot \frac{100}{64} \right\rceil 37 = \left\lfloor \mathbf{64} \cdot \frac{100}{64} \right\rceil 37 + \left\lfloor \mathbf{73} \cdot \frac{100}{64} \right\rceil 37\,, \tag{22}$$

from which follows (21). For the other nucleon regularities (Table 3), it is less obvious to establish similar relations, but their presence can be seen through the average nucleon numbers which are approximately $10\Psi^2 \approx 137$ for $\mathcal{A}^\times$ and $100\Psi \approx 370$ for $C$, so that an approximate value of the aggregate nucleon numbers is $10(\Psi^2 + 10\Psi) = 10\Psi^3 \approx 507$ in SGC (notice that $\left|\mathcal{A}^\times\right| = |C| = 61$).

Shcherbak [4] showed that each canonical amino acid has in a main-chain (a standard block) exactly $74 = 2 \cdot 37$ nucleons, as well as that the total nucleon numbers of the main-chains, the side-chains and their sum in $4*(\mathcal{A})$ are the squares of the first three Pythagorean numbers multiplied by 37 [4]. Actually, Pythagorean numbers can be also obtain from $\Psi$ and $\psi$ as

$$\left[\Psi\frac{10}{4}\right]+\left[(\Psi+\psi)\frac{10}{4}\right]=\left[\Psi\psi\frac{10}{4}\right] \quad \Rightarrow \quad 3^2+4^2=5^2. \tag{23}$$

However, all presented regularities require further research and reconsideration, which potentially could be directed to the similar analyzes of protons/electrons or molecular masses, which is supported by the fact that the total *molecular mass* of $\mathcal{A}$ is $2738.04 = 2\cdot 37^2 + 0.04$ [36], unlike for the nucleon numbers of $\mathcal{A}$ when is $2735 = 2\cdot 37^2 - 3 = 200\cdot 13.7 - 5$. Further dual scaling analysis could also be carried out for the (probably slightly) changed parameters, such as the scaled values of an exact value of the fine structure constant. Generally, these preliminary results should be subjected to detailed and rigorous considerations using some statistical analyses, and consequently reconsider the conclusions, what is also in the plan.

In the context of here presented results, most generally can be said that the nucleon regularities support a dual nature of genetic code – genome/protein, codons/amino acids, free molecules/molecule residues, fourfold/twofold degenerate code halves, S/W properties, Y/R properties, M/K invariance in a way that these dual domains are generally determined by dual scaling constants, 37 and 13.7. These regularities can be partly resulted from some underlying dualities, and even the ultimate quantum-classic duality, where the last can give a direct and general answer to the origin of nucleon regularities. The fact that scaling constants 37 and 13.7 result from the discrete inverse mapping, supports such hypothesis, while the closeness of value $10\Psi^2$ to the fine structure constant could potentially indicate on the origin of this constant from some (holographic) duality. A possible nonlinear nature of the quantum-classic duality in mesoscopic domain could facilitate the emergence of complex structures, what is characteristic of the living organisms and thus indicates that their appearances at the mesoscopic scales are nontrivial. Such mesoscopic nonlinear regime also can contribute to the sensitivity of the system to the boundary conditions, initial conditions, parameter values etc. Finally, the last favors the existence environmental-dependent stereochemical mechanism, which inherently could act in some degree as the error-minimization mechanism, throughout the entire period of the genetic code emergence, what supports Knight's scenario of complementary evolutionary forces [2], as well as the theories of a mineral-mediated origin of the genetic code [7,8].

## 6. CONCLUSION

In this paper is considered some potential underlying selective mechanisms involved in the origin and evolution of the genetic code. Starting from the assumption that in the genetic code, particularly SGC, the various arithmetical regularities of the nucleon numbers of amino acids based on the decimal number 37 could be so far unknown relic of stereochemical mechanism or even of the error-minimization mechanism, here is analysis extended on some new division of genetic code related to its *p*-adic model and for the additional parameter – the number 13.7. Also analysis is extended on the codons, how separately so conjointly with the amino acids in the form of the aggregate nucleon numbers. The results are also considered in a context of some selfsimilar and scaling properties of 37 and 13.7. Some main results and conclusions will be given in sequel.

The use of Rumer's canonical nucleobase order for the *p*-adic description of the genetic code in the form $\{A,U,G,C\} = \{1,2,3,4\}$ results in the closest 2-adic codon-amino acid assignment.

Generally, SGC shows systematize determination by 37 and 13.7 on the level of codons and amino acids, how separately so conjointly. All so far investigated nucleon regulaties in VMC shows that they are originate from SGC, i.e. all changes in VMC have resulted negatively in terms of the nucleon regularities, what suggests that the selective mechanism of SGC origin and VMC evolution are different to some significant degree and hence promotes the stereochemical theory.

The replacement $U^{RNA} \leftrightarrow T^{DNA}$ is such that, in terms of the nucleon numbers, makes RNA nucleobase space better tuned for a codon degeneracy, while DNA nucleobase space better tuned for a strain pairing. The nucleon number divisibility for both RNA and DNA nucleobase triplets – codons, shows no correspondence for the total codon spaces, but only for the reduced codon spaces of SGC by 3 termination codons, what is again an argument for the sterochemical theory which is based on affinity between the codons and the amino acids, and not between the codons and the release factors. The code degeneracy pattern reflects the most faithful image of 37 and 13.7 divisibility, not only because almost strict correspondence, but also because all regularities are coherently realized, how separately at the level of codons and amino acids, so consequently together and even on the disjoint sets YNN and RNN. Regularities are also shown for the division according to the first base, Y/R and S/W nucleobase properties for the aggregate nucleon numbers. Within the framework of the discussed methodology, SGC shows general tendency to be organized as the two-leveled hierarchical structure where each of them is tuned separately by 37 and 13.7 – typically, the free amino acids, the fourfold degenerate code half and SNN are better tuned by 37, while oppositely the codons, the amino acids residues, the two fold degenerate code half and WNN are better tuned by 13.7. The average nucleon numbers of the amino acid sets are approximately $10 \cdot 13.7 \approx 10 \cdot 3.7^2$, while the codon set is $10 \cdot 37$, what overall results an approximate average value of the aggregate nucleon numbers $10 \cdot 37^3$ in SGC, as well as in VMC. Some nucleon divisibility regularities are founded also for the amino acid residues and their ionized/protonated forms, which are dominantly related to 13.7, oppositely to the free amino acid which are dominantly related to 37.

The presented nucleon regularities support a multiple dual nature of the genetic code and open the possibility that they are ultimately emerged from the quantum-classical duality, what can potentially and to some degree give a direct and general answer to the origin of nucleon regularities. Some main potential consequences would be a nontrival emergence of the living organisms at the mesoscopic scales, the existence of environmental-dependent stereochemical mechanism and a mineral-mediated origin of the genetic code.

## REFERENCES

1. SZATHMÁRY, E.: *The origin of the genetic code: amino acids as cofactors in an RNA world*. Trends Genet. **15** (1999) 223-229.
2. KNIGHT, R.D., FREELAND, S.J., LANDWEBER, L.F.: *Selection, history and chemistry: the three faces of the genetic code*. Trends Biochem. Sci. **24** (1999) 241-247.

3. KNIGHT, R.D., FREELAND, S.J., LANDWEBER, L.F.: *Rewiring the keyboard: evolvability of the genetic code*. Nat. Rev. **2** (2001) 49-58.
4. SHCHERBAK, V.I.: *Sixty-four triplets and 20 canonical amino acid of the genetic code: the arithmetical regularities. Part II*. J. Theor. Biol. **166** (1994) 475-477.
5. MIŠIĆ, N.Ž.: *From genetic code toward spacetime geometry*. In Proceedings of the 2nd International Conference on Theoretical Approaches to BioInformation Systems, Belgrade (2014) 101-123.
6. DRAGOVICH, B.: *p-Adic structure of the genetic code*. Neuroquantology **9** (2011) 716-727.
7. CLEAVES, H.J., SCOTT, A.M., HILL, F.C., LESZCZYNSKI, J., SAHAI, N., HAZEN, R.: *Mineral-organic interfacial processes: potential roles in the origins of life*. Chem. Soc. Rev. **41** (2012) 5502-5525.
8. FERRIS, J.P.: *Mineral catalysis and prebiotic synthesis: montmorillonite-catalyzed formation of RNA.* Elements **1** (2005) 145-149.
9. OLSEN, G.J., WOESE, C.R.: *Lessons from an archaeal genome: What are we learning from Methanococcus jannaschii*. Trends Genet. **12** (1996) 377–379.
10. WOESE, C.R.: *On the evolution of cells*. Proc. Natl. Acad. Sci. USA **99** (2002) 8742-8747.
11. RUMER, Y.B.: *About systematization of the codons in the genetic code*. Proc. Acad. Sci. USSR **167** (1966) 1393-1394.
12. RUMER, Y.B.: *Systematization of the codons in the genetic code*. Proc. Acad. Sci. USSR **183** (1968) 225-226.
13. YANG, M.M., COLEMAN, W.J., YOUVAN, D.C.: *Genetic coding algorithms for engineering membrane proteins*. In: Reaction Centers of Photosynthetic Bacteria (Michel-Beyerle, M., ed.) Springer-Verlag, Berlin (1990) 209-218.
14. WONG, J.T.: *A co-evolution theory of the genetic code*. Proc. Natl. Acad. Sci. USA **72** (1975) 1909-1912.
15. TAYLOR, F.J.R., COATES, D.: *The code within the codons*. Biosystems **22** (1989) 177-87.
16. WETZEL, R.: *Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code*. J. Mol. Evol. **40** (1995) 545-550.
17. SONNEBORN, T.M.: *Degeneracy of the genetic code: extent, nature, and genetic implications*. In: Evolving Genes and Proteins (Bryson, V. and Vogel, H. J., eds.). Academic Press, New York (1965) 377-397.
18. WOESE, C.R.: *On the evolution of the genetic code*. Proc. Natl. Acad. Sci. USA **54** (1965) 1546-1552.
19. HAIG, D., HURST, L.D.: *A quantitative measure of error minimization in the genetic code*. J. Mol. Evol. **33** (1991) 412-417.
20. CHECHETKIN, V.R.: *Block structure and stability of the genetic code*. J. Theor. Biol. **222** (2003) 177-188.
21. KING, J.L., JUKES, T.H.: *Non-Darwinian evolution*. Science **164** (1969) 788-798.
22. HASEGAWA, M. AND MIYATA, T.: *On the asymmetry of the amino acid code table*. Orig. Life. **10** (1980) 265-270.
23. DOWNES, A.M., RICHARDSON, B.J.: *Relationships between genomic base content and distribution of mass in coded proteins*. J. Mol. Evol. **55** (2002) 476-490.
24. SHCHERBAK, V.I.: *Arithmetic inside the universal genetic code*. Biosystems **70** (2003) 187-209.
25. DI GIULIO, M.: *The origin of the genetic code: theories and their relationships, a review.* Biosystems **80** (2005) 175-184.
26. KOONIN, E.V., NOVOZHILOV, A.S.: *Origin and evolution of the genetic code: the universal enigma*. IUBMB Life **61** (2009) 99-111.
27. SENGUPTA, S., HIGGS, P.G.: *Pathways of genetic code evolution in ancient and modern organisms.* J. Mol. Evol. **80** (2015) 229-243.
28. BLIZARD, W.D.: *Multiset theory*. Notre Dame J. Formal Logic **30** (1988) 36-66.

29. DRAGOVICH, B., KHRENNIKOV, A.YU., KOZYREV, S.V., VOLOVICH, I.V.: *On p-adic mathematical physics*. *p*-Adic Numbers Ultrametric Anal. Appl. **1** (2009) 1-17.
30. KOBLITZ, N.: *p-adic Numbers, p-adic Analysis, and Zeta-Functions* (*2nd ed.*). Springer-Verlag, New York, 1984.
31. ROBERT, A.M.: *A Course in p-adic Analysis*. Springer-Verlag, New York, 2000.
32. DRAGOVICH, B., KHRENNIKOV, A.YU., MIŠIĆ, N.Ž.: *Ultrametrics in the genetic code and the genome*. Appl. Math. Comput. (2016) (submitted).
33. MARCK. C., GROSJEAN, H.: *tRNomics: Analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features*. RNA **8** (2002) 1189-1232.
34. TRIFONOV, E.N.: *The triplet code from first principles*. J. Biomol. Struct. Dyn. **22** (2004) 1-11.
35. VERKHOVOD, A.B.: *Alphanumerical divisions of the universal genetic code: new divisions reveal new balances*. J. Theor. Biol. **170** (1994) 327-330.
36. RAKOČEVIĆ, M.M.: *A harmonic structure of the genetic code*. J. Theor. Biol. **229** (2004) 221-234.
37. MIŠIĆ N.Ž.: *The self-similar numbers as a special case of cyclic numbers and their relation to the cyclic (genetic) codes*. Proceeding of 10th Symposium on Neural Networks and Applications (2010) 97-102.
38. MIŠIĆ, N.Ž.: *Nested numeric/geometric/arithmetic properties of Shcherbak's prime quantum 037 as a base of (biological) coding/computing*. Neuroquantology **9** (2011) 702-715.
39. SNYDER, W.M.: *Factoring Repunits*. Am. Math. Monthly **89** (1982) 462-466.

# STANDRADNI GENETSKI KOD: *P*-ADIČKO MODELIRANJE, NUKLEONSKE RAVNOTEŽE I SAMOSLIČNOST

*Rad predstavlja preliminarne rezultate i zaključke na jedno od fundamentalnih pitanja genetskog koda u vezi sa osnovnim selektivnim mehanizmima uključenim u njegov nastanak i evoluciju, a naročito njihova hipotetička različita priroda, prvobitno razmatrana u [1,2,3]. Uveden je novi pristup zasnovan na poznatim aritmetičkim pravilnostima unutar genetskog koda, odredjenim različitim ravnotežama broja nukleona aminokiselina i njihovom deljivošću decimalnim brojem 37 [4]. Kao parametar sistematizacije genetskog koda uveden je združeni broj nukleona aminokiseline i srodnog kodona, dok je test deljivosti sproveden ne samo brojem 37, već i 13,7, kao konstantom samosličnosti decimalnog skaliranja [5]. Relevantne nukleonske sume su odredjene za najistaknutije podele standardnog genetskog koda (SGK) prema p-adičkom modelu mitohondrijalnog koda kičmenjaka (MKK) [6]. Takodje su analizirani obrasci deljivosti nukleonskih suma brojevima 37 i 13,7 za RNK i DNK kodonski prostor, kao i za aminokiselinski prostor. Dobijeni rezultati, pre svega viša deljivost nukleonskih suma brojevima 37 i 13,7 u SGK nego u MKK, kao i korespondencija izmedju obrasca deljivosti nukleonskih brojeva u RNK kodonskom prostoru i aminokiselinskom prostoru SGK-a, kako odvojeno tako i zajedno, sa obrascem degeneracije koda, sugerišu neke zaključke: podržava hipotezu [1,2,3,7] da selektivne pokretačke sile koje deluju tokom nastanka (drevna faza) i evolucije (moderna faza) genetskog koda jesu različite, ukazuje na postojanje stereohemijskog mehanizma zavisnog od okruženja tokom celog perioda nastanka genetskog koda i podržava poreklo genetskog koda koje je posredovano mineralima [7,8].*

**Ključne reči:** *genetski kod, poreklo i evolucija, stereohemijski mehanizam, nukleonske ravnoteže, samosličnost*