BioSystems 185 (2019) 104017

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Review article

p-Adic hierarchical properties of the genetic code

Branko Dragovich[a,b,*], Nataša Ž. Mišić[c]

[a] Institute of Physics, University of Belgrade, Belgrade, Serbia
[b] Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
[c] Research and Development Institute Lola Ltd, Kneza Višeslava 70a, Belgrade, Serbia

ARTICLE INFO

Keywords:
Genetic code
p-Adic distance
Ultrametric codon tree
Amino acids
Genetic language
p-Adic network
Bioinformation

ABSTRACT

In this article, we consider *p*-adic modeling of the standard genetic code and the vertebrate mitochondrial one. To this end, we use 5-adic and 2-adic distance as a mathematical tool to describe closeness (nearness, similarity) between codons as elements of a bioinformation space. Codons which are simultaneously at the smallest 5-adic and 2-adic distance code the same (or similar) amino acid or stop signal. The set of codons is presented as an ultrametric tree as well as a fractal and *p*-adic network. It is shown that genetic code can be treated as sequential translation between genetic languages. This *p*-adic approach gives possibility to be applied to sequences of nucleotides of an arbitrary finite length. We present an overview of published and some new results on various *p*-adic properties of the genetic code.

1. Introduction

The genetic code plays a central role in all living organisms. It connects codons in genes and amino acids in proteins, as well as it determines codons responsible for stop signal in synthesis of proteins. From a mathematical point of view, the genetic code is a map from the set of 64 elements (which are codons) onto 21 elements (which are 20 amino acids and 1 stop signal). In other words, it is a general and a firm experimental fact that tells us which concrete codons code any of amino acids, and which codons code stop (terminal) signal.

The genetic code is a result of the origin and early evolution of life at our planet that happened almost four billion years ago. There are no concrete facts from that early period of time and it gives rise to very different hypotheses. So stereochemical theory and the coevolution theory are strictly based on some physicochemical processes. However, many processes at primitive Earth were very complex and stochastic. Hence the very beginning of formation of the genetic code should be treated with ambiguity (arbitrariness) which was gradually under reduction until it became "frozen" (Crick, 1968). For a review on these approaches to the origin, early evolution and meaning of the genetic code, see Barbieri (2015, 2018, 2019).

The genetic code is experimentally deciphered in the mid of 1960s (Bernfield and Nirenberg, 1965) and is usually presented by the corresponding table, where it is explicitly written the relation between codons and amino acids with stop signal. Hence, one can say that we experimentally know what the genetic code is. However, we cannot say that there is its complete theoretical description and understanding. Namely, there are numerous questions that still have not satisfactory answers. The main question is: Why the genetic code is just such as it is? For example, there are theoretically about 1.5×10^{84} (Dragovich et al., 2017b) possibilities to connect codons and amino acids, but in living organisms have been found only a few dozens. Another important problem is finding of theoretical background for structure of *degeneracy* of the genetic code. Codon degeneracy consists in the fact that almost all amino acids are coded by more than one codon. For example, in the vertebrate mitochondrial code an amino acid is coded by two, four or six codons, see Table 3 in this paper.

It is worth noting difference between description and understanding. It can be illustrated by situation with Quantum Mechanics. Namely, there are a few theoretical formalisms by which we can describe quantum phenomena with predictions in quite good agreement with experimental data. However, we do not understand what really happens inside quantum systems, as we understand it in the case of classical ones. We should keep in mind this distinction between description and understanding when want theoretically consider the genetic code. Hence, theoretical priority should be adequate description and then possible understanding.

The starting point in theoretical understanding of any phenomenon should be construction of an adequate model. A valuable model should at least contain a correct description of some essential characteristics of the subject under consideration. In modeling the genetic code there are many, and rather different, approaches that usually depend on the scientific background of the scientist who is doing it. So, there are physical, mathematical, chemical, biological and some other approaches. A brief review of some theoretical modelings of the genetic code we present in Section 3.

The main subject which we consider in this article is modeling of the genetic code by using *p*-adic distance as a simple and adequate

* Corresponding author at: Institute of Physics, University of Belgrade, Belgrade, Serbia.
E-mail addresses: dragovich@ipb.ac.rs (B. Dragovich), nmisic@rcub.bg.ac.rs (N.Ž. Mišić).

https://doi.org/10.1016/j.biosystems.2019.104017
Received 22 May 2019; Received in revised form 10 August 2019; Accepted 13 August 2019
Available online 18 August 2019

mathematical tool to describe its many properties. This approach was proposed in 2006 (Dragovich and Dragovich, 2006). This model has been further considered and developed mainly in Dragovich and Dragovich (2007, 2009, 2010), Dragovich (2012, 2016), Dragovich et al. (2017b), Dragovich and Mišić (2018) and Mišić (2016). The basic idea is as follows. Two codons which code the same amino acid are closer in the information sense than those codons that code different amino acids. This closeness (nearness, similarity) is in an ultrametric space which elements are codons. The natural way to describe such ultrametric space is by introducing relevant p-adic space of codons. In Dragovich's approach, relevant 5-adic space was constructed and p-adic distance between codons was considered for $p = 5$ and $p = 2$. As a result one obtains that two codons which are simultaneously closest under 5-adic and 2-adic distance code the same amino acid or stop signal. This strongly holds in the case of the vertebrate mitochondrial code and situation is slightly different for the standard and other genetic codes. This model very well describes the degeneracy of the genetic code and also some other its properties. Model of the genetic code on the dyadic plane is presented in Khrennikov and Kozyrev (2007).

In this paper, we review the main published results and present our new recent achievements in investigation of p-adic genetic code properties. Section 2 contains an introduction to p-adic distance, because it is basic mathematical tool in this analysis of the genetic code. In Section 3 we present some basic facts on the gene expression and a brief review of the genetic code modeling. To p-adic modeling of the genetic code is devoted Section 4, where various aspects of the standard and vertebrate mitochondrial genetic codes are considered. In particular, it is shown that the genetic code can be viewed as translation between four genetic languages, as well as an ultrametric tree, a fractal and a p-adic network. Proposition to apply p-adic distance to study similarity between other biomolecular sequences is subject of Section 5, which contains also some concluding remarks.

2. p-Adic distance

It is well known that the usual distance is an adequate mathematical tool for the description of positions in our ordinary space. However, it is not sufficiently known that p-adic distance can be used to measure positions in a bioinformation space. In particular, p-adic distance serves as an adequate mathematical tool to describe closeness (nearness) of information in the genetic coding. For example, codons which are at the smallest p-adic distance code the same amino acid. We are going now to present p-adic distance in a way to be comprehensive to a wide audience interested in theoretical aspects of the genetic code.

Note that in the word "p-adic", p denotes a prime number. Recall that a prime number is a natural number greater than 1 that is divisible only by itself and 1. There are infinitely many prime numbers and the first 5 of them are: 2, 3, 5, 7, 11. Any natural number greater than 1, is either a prime or composite number, which can be presented (up to the order) as a unique product of primes. For example 4 and 6 are composite numbers, because $4 = 2 \cdot 2$, $6 = 2 \cdot 3$. Prime numbers play a fundamental role in number theory and many other parts of mathematics. In the sequel, we are mainly interested in p-adic aspects of integer numbers.

An important notion is p-adic norm, which can be defined in the following way. Any integer number $m \neq 0$, with respect to a given prime p, can be presented as $m = p^k a$, where $k \in \{0, 1, 2, 3, \ldots\}$ and a in an integer not divisible by p. Then, by definition, p-adic norm of $m \neq 0$ (denoted by $|m|_p$) is $|m|_p = p^{-k}$. If $m = 0$, then by definition $|0|_p = 0$. One can illustrate p-adic norm by the following example:

$$|360|_p = |2^3 \times 3^2 \times 5|_p = \begin{cases} \frac{1}{8} & \text{if } p = 2, \\ \frac{1}{9} & \text{if } p = 3, \\ \frac{1}{5} & \text{if } p = 5, \\ 1 & \text{if } p \geq 7. \end{cases} \tag{1}$$

Now one can define p-adic distance between any two integers x and y as

$$d_p(x, y) = |x - y|_p. \tag{2}$$

For example, let $x = 63$ and $y = 3$. Then

$$d_p(63, 3) = |63 - 3|_p = |60|_p = |2^2 \times 3 \times 5|_p = \begin{cases} \frac{1}{4} & \text{if } p = 2, \\ \frac{1}{3} & \text{if } p = 3, \\ \frac{1}{5} & \text{if } p = 5, \\ 1 & \text{if } p \geq 7. \end{cases} \tag{3}$$

From definition (2) and example (3), one can easily conclude that: (a) for fixed numbers x and y, p-adic distance depends on p, (b) for infinitely many primes p, distance is the same and equal to 1, (c) p-adic distance has discrete values, (d) inequality $d_p(x, y) \leq 1$ holds for any integers x and y and any prime p.

p-Adic norm and distance are very different from the well known usual absolute value (denoted by $|x|$) and related distance $d(x, y) = |x - y|$. Recall that absolute value of number x is

$$|x| = \begin{cases} x & \text{if } x \text{ is positive}, \\ -x & \text{if } x \text{ is negative}, \\ 0 & \text{if } x = 0. \end{cases} \tag{4}$$

However, there is a nice connection between p-adic norms and absolute value in the form of their product, i.e.

$$|x| \prod_p |x|_p = 1, \tag{5}$$

where x is any integer different from 0. Formula (5) is a consequence of the fact that a non-zero integer x can be presented as product $x = \pm p_1^{k_1} p_2^{k_2} \ldots p_n^{k_n}$, where $p_1, p_2, \ldots, p_n$ is a finite set of some prime numbers, and $k_1, k_2, \ldots, k_n$ are some natural numbers. Then $\prod_p |x|_p = p_1^{-k_1} p_2^{-k_2} \ldots p_n^{-k_n}$, while ordinary absolute value $|x| = p_1^{k_1} p_2^{k_2} \ldots p_n^{k_n}$. Product formula (5) also holds when x is a non-zero rational number.

p-Adic distance is the most significant example of ultrametrics. The ultrametric space is a set M endowed with a distance d that has the following properties:

$$\begin{aligned} &(i) \quad d(x, y) \geq 0, \ \ d(x, y) = 0 \text{ if and only if } x = y, \\ &\text{(ii)} \quad d(x, y) = d(y, x), \\ &\text{(iii)} \quad d(x, y) \leq \max\{d(x, z), d(z, y)\}, \end{aligned} \tag{6}$$

where $x, y, z \in M$. Property (*iii*) is usually called the strong triangle inequality, but it is also known as ultrametric or non-Archimedean inequality. p-Adic distance defined in (2) satisfies all properties in (6). On some applications of ultrametrics in physics and biology, one can refer to review (Rammal et al., 1986). Ultrametrics is adequate mathematical tool for description of the nested structure in hierarchical systems.

According to (2) and properties of p-adic norm, it follows that two integers are closer as their difference is more divisible by prime number p. It can be also well seen presenting numbers by expansion in base p. Namely, let

$$\begin{aligned} x &= x_0 + x_1 p + x_2 p^2 + \cdots + x_k p^k \equiv x_0 x_1 x_2 \ldots x_k, \\ y &= y_0 + y_1 p + y_2 p^2 + \cdots + y_k p^k \equiv y_0 y_1 y_2 \ldots y_k, \end{aligned} \tag{7}$$

where $x_i \in \{0, 1, \ldots, p-1\}$ and $y_i \in \{0, 1, \ldots, p-1\}$ are related digits. Then, p-adic distance between x and y from (7) is

$$d_p(x,y) = |x-y|_p$$
$$= \begin{cases} 1 & \text{if } x_0 \neq y_0, \\ \frac{1}{p} & \text{if } x_0 = y_0,\ x_1 \neq y_1, \\ \frac{1}{p^2} & \text{if } x_0 = y_0,\ x_1 = y_1,\ x_2 \neq y_2, \\ \text{---} \quad \text{---} \quad \text{---} \quad \text{---}, \\ \frac{1}{p^k} & \text{if } x_0 = y_0, \ldots, x_{k-1} = y_{k-1},\ x_k \neq y_k. \end{cases} \tag{8}$$

Expressions (8) nicely illustrate how p-adic distance depends on digits. If first digits are different then distance is maximal and does not depend on values of the other digits. Two integers are closer as they have more equal digits looking from the beginning.

Integers endowed by ordinary absolute value are real integers, but integers endowed by p-adic norm (also named p-adic absolute value) are called p-adic integers. Note difference in presenting real and p-adic integers by their digits, which is in opposite way, for example:

$$x = x_k p^k + \cdots + x_1 p + x_0 \equiv x_k \ldots x_1 x_0 \quad \text{(real)},$$
$$x = x_0 + x_1 p + \cdots + x_k p^k \equiv x_0 x_1 \ldots x_k \quad (p\text{−adic}).$$

This is a consequence of the fact that in real case term $x_k p^k$ is the largest, while in p-adic case is the smallest, and vice versa. What is “large” or “small” depends on the applied norm (ordinary or p-adic absolute value).

p-Adic numbers, which include integers, play an important role in p-adic analysis (Schikhof, 1985; Robert, 2000). p-Adic numbers are invented by K. Hensel in 1897. It is worth mentioning that p-adic analysis has applications in modeling systems with hierarchical structure and it is known as *p-adic mathematical physics*, for some reviews see (Brekke and Freund, 1993; Vladimirov et al., 1994; Dragovich et al., 2009, 2017a). In p-adic modeling of the genetic code it is mainly sufficient to use the relevant distance between integers.

3. Genetic code and its modeling

To have this paper self-contained, we give here a very brief review of some main facts about gene expression, the genetic code and its modeling.

3.1. Along gene expression

All biomolecular components and processes related to the genetic code are contained inside living cells.

Information on the primary structure of proteins and processes regulation is stored in DNA (deoxyribonucleic acid). DNA is a macromolecule composed of two polynucleotide strands mutually coiled into a double helix. A nucleotide is composed of a nitrogenous base, a sugar and a phosphate group. DNA contains four bases: adenine (A), cytosine (C), guanine (G) and thymine (T). Adenine and guanine are purines that are composed of two carbon-nitrogen rings. Cytosine and thymine are pyrimidines, which contain one carbon-nitrogen ring. In an informational sense, nucleobases and nucleotides have the same meaning. Bases in one and the other chain are connected in a complementary way making base pairs A-T and C-G. In the human DNA, there are about 3×10^9 base pairs and about 20,000 protein-coding genes. In addition to protein-coding genes, DNA also contains a noncoding part, which is related to instructions for regulation of gene expression and other cell processes. There are two processes to extract genetic information from DNA: replication and transcription. By replication DNA duplicates giving two DNAs with the same genetic information as the original one.

Ribonucleic acid (RNA) is usually a single polynucleotide chain. There are many kinds of RNA, in particular microRNA, that play various roles in processes from transcription to protein synthesis. As a result of gene transcription from DNA, one obtains messenger RNA (mRNA), so that bases A, C, G, T are respectively transcribed to their complements U, G, C, A, where U denotes uracil (which is a pyrimidine). This mRNA, copied from DNA, contains introns and exons, and during the splicing process introns are moved and one gets mature mRNA containing only exons. In the splicing, process exons can be joined by different ways giving various possibilities, and as a result from one gene one can obtain many proteins.

Mature mRNA contains a sequence of trinucleotides, i.e. codons, that are related to amino acids or a stop signal. As adaptors between codons and amino acids serve transfer (transport) RNA (tRNA) and aminoacyl-tRNA synthetases. A transfer RNA is 75–90 nucleotides long chain that has cloverleaf structure and contains an anticodon. An aminoacyl-tRNA synthetase is a protein that attaches appropriate amino acid onto corresponding tRNA according wobble rules. To each of 20 amino acids there is only one aminoacyl-tRNA synthetase.

Protein synthesis performs in ribosomes, which are complex macromolecular machines composed of ribosomal RNAs and ribosomal proteins. A ribosome has two subunits: smaller subunit reads sequence of codons in messenger RNA, while larger subunit joins amino acids into polypeptide chain in the same order as codons are in mRNA. Translation of codons into amino acids is strictly according to the genetic code.

Proteins (Finkelstein and Ptitsyn, 2002) are long polypeptides which primary structure is a chain of amino acids determined by sequence of codons in the mRNA. To improve functionality, proteins may have also post-translation modification. Final three-dimensional structure is a result of folding and this structure is closely related to their functions. Performing various functions, proteins are essential constituents of all living cells.

According to the aforementioned machinery along gene expression, the primary role of the genetic code is to adequately translate genetic information, stored in a protein-coding sequence of triplets composed of four kinds of nucleotides in DNA, to the related sequence of 20 canonical amino acids. This fact suggests considering the genetic code as a translation from the four-letter language of DNA to the 20-letter language of proteins. This language aspect will be elaborated later in the present paper.

3.2. A brief review of genetic code modeling

First theoretical considerations of the genetic code were begun soon after the discovery of the helicoidal structure of DNA by Watson and Crick (1953). The first model was constructed by G. Gamow, whose scientific background was in physics, and is known as *diamond code* (Gamow, 1954). Gamow's diamond code and *comma-free code* (Crick et al., 1957) were “brilliant ideas that turned out to be utterly wrong” because biological aspects were not taken into account (for a review of this first stage of the genetic code modeling, see Hayes, 1998). However, Gamow's idea that genetic material should be coded by triplets of nucleotides was shown to be correct.

More realistic models began to appear during the experimental discovery of codons to amino acids assignment, what started by M. W. Nirenberg and J. H. Matthaei in 1961 and mostly finished in 1965 (Bernfield and Nirenberg, 1965). Experimental deciphering of the genetic code in the mid of 1960s has given rise to many theoretical attempts to describe and understand its properties. Despite a lot of published papers, still there is no theoretical model that completely describes all properties of the genetic code and investigations last until today. In the sequel of this section we shall mention only several models, mainly to illustrate variety of approaches.

Rumer (1966, 1968) noted that there are 32 codons grouped into 8 quadruplets for which nucleotide at the third position is irrelevant for coding. As a consequence, he introduced notions of 8 strong (CC, AC, UC, GC, CU, GU, CG, GG) and 8 week (CA, AA, UA, GA, AU, UU, AG, UG) roots of codons, and according to “strength” classified nucleotides (C – very strong, G – strong, U = T – weak, and A – very weak). He also

considered symmetric transformations between bases and between codons. For a brief review on Rumer's life and work on the genetic code, see (Fimmel and Strüngmann, 2016).

Crick (1966) proposed wobble hypothesis for codon-anticodon pairing on the third base to explain the degeneracy of the genetic code. This will be presented in some details in the next section. Crick (1968) considered origin of the genetic code and concluded that after some developments at the beginning it was frozen.

C. R. Woese devoted a series of papers to the genetic code and its evolution, wishing "to know the mechanisms giving rise to the particular assignments" (Woese, 1965). His main contribution to the evolution of life is discovery of the third domain, i.e. *Archea* (Woese and Fox, 1977).

Investigation of the genetic code from the point of view of the Grey code is presented in Swanson (1984), where amino acids are divided into four groups (small, large, external, internal).

In 1990s appeared a series of papers that contained algebraic approach to consider structure of the genetic code. A motivation for employment of algebraic methods, in fact some Lie groups and algebras, comes from high energy theoretical physics. This research was initiated by paper (Hornos and Hornos, 1993), where 64-dimensional irreducible representation of the symplectic $Sp(6)$ group and its irreducible representations of subalgebras arising in a chain of symmetry breaking are compared with degeneracy of the genetic code and obtained interesting approximative result. For similar supersymmetric models of the genetic code one can see Bashford et al. (1997) and Forger and Sachse (2000). Another model based on a quantum algebra was investigated, see Frappat et al. (2001) and references therein.

Shcherbak (2003, 2008) found various arithmetic regularities inside the genetic code determined by the masses of its constituents (the amino acids and nucleobases), primarily for a code degeneracy pattern in the form of aforementioned Rumer's division (Shcherbak, 1993, 1994). Further analysis of the mass distribution confirmed the existence of mass balances at the protein level (Downes and Richardson, 2002). These studies have motivated new mathematical approaches (Rakocevic, 2004) and deepened the "enigma" of the genetic code origin related to the underlying selective mechanisms involved in its shaping (Mišić, 2014, 2016).

p-Adic approach to modeling the genetic code was proposed in 2006 (Dragovich and Dragovich, 2006) and further developed mainly in Dragovich and Dragovich (2009), Dragovich and Dragovich (2010), and Dragovich et al. (2017b).

As recent review on noise immunity, frame shift problem, circular codes and some symmetries, we refer to Fimmel and Strüngmann (2018).

To understand the genetic code, in addition to modeling modern properties, it is also necessary to know its origin and evolution. Origin and evolution of life and the genetic code are interrelated. For these investigations we refer to recent reviews (Koonin and Novozhilov, 2009; Barbieri, 2018; Kun and Radványi, 2018) and references therein. Origin and early evolution of the genetic code was also conjectured within p-adic approach (Dragovich and Dragovich, 2010), so that there were three steps in generating codons (single nucleotides → dinucleotides → trinucleotides).

4. p-Adic modeling of the genetic code

4.1. Genetic code as translation of genetic languages

It is useful to consider the genetic code as a translation between a few forms of genetic language. To this end, it is worth to recall some relevant notions from the theory of formal languages.

An *alphabet* is usually a finite set of elements which are called *letters*, e.g. it can be denoted as $\mathcal{A} = \{a, b, c, ...\}$. A *word* over alphabet $\mathcal{A}$ is any finite string (sequence) of letters, i.e. words of length n are elements of $\mathcal{A}^n = \{x_1x_2...x_n : x_i \in \mathcal{A}\}$. Let the set of all words over alphabet $\mathcal{A}$ be denoted as $\mathcal{A}^*$. Then a language $\mathcal{L}$ over alphabet $\mathcal{A}$ is a subset of $\mathcal{A}^*$, i.e. $\mathcal{L}$ is a subset of all words over alphabet $\mathcal{A}$.

According to the previous paragraph one can introduce four kinds of biomolecular languages with related alphabets. These four alphabets

$$\mathcal{A}_1 = \{A, C, G, T\}, \tag{9}$$

$$\mathcal{A}_2 = \{A, C, G, U\}, \tag{10}$$

$$\mathcal{A}_3 = \{A, C, G, U, I\}, \tag{11}$$

$$\mathcal{A}_4 = \{M_1, M_2, ..., M_{20}\}, \tag{12}$$

produce the corresponding languages

$$\mathcal{L}_1 = \{N_1N_2N_3 : N_i \in \mathcal{A}_1\}, \tag{13}$$

$$\mathcal{L}_2 = \{N_1N_2N_3 : N_i \in \mathcal{A}_2\}, \tag{14}$$

$$\mathcal{L}_3 \subset \{N_1N_2N_3 : N_i \in \mathcal{A}_3\}, \tag{15}$$

$$\mathcal{L}_4 = \{\text{a set of proteins}\}, \tag{16}$$

where A, C, G, T, U are standard nucleotides, I is inosine, and M_1, M_2, …, M_{20} are twenty canonical amino acids. Inosine is nucleoside whose counterpart of base is hypoxanthine, which is a purine derivative. Hence alphabets $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ altogether contain three pyrimidines (C, T, U) and three purines (A, G, I). Inosine is a constituent of some tRNA and serves in the genetic code translation by wobble base pairing.

In languages $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ there are three-letter words called codons. Languages $\mathcal{L}_1$ and $\mathcal{L}_2$ contain 64 words (codons), and all of them are employed in the genetic code. $\mathcal{L}_3$ contains 125 words, however only some of them are incorporated in tRNA according wobble pairing. These languages are illustrated in Fig. 1.

The first step in gene expression is transcription with sequent splicing of a gene from DNA to related mature mRNA. This corresponds to the translation of genetic information written in language $\mathcal{L}_1$ to language $\mathcal{L}_2$. In this process codons in the gene are translated to their anticodons according to complementary base pairing, i.e.

$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$
C	G	C, I
A	U	A, G, I
T	A	U, I
G	C	G, I

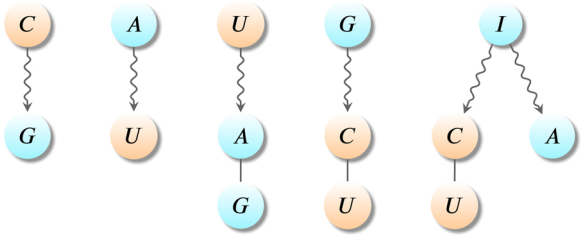

Fig. 1. The figure contains three levels. At the top level are the first three languages in the translation of genetic information. The middle level contains translations of nucleotides in translation of languages $\mathcal{L}_1 \to \mathcal{L}_2 \to \mathcal{L}_3$. Translation of letters from $\mathcal{L}_2$ to $\mathcal{L}_3$ is not one-to-one. At the bottom level is an illustration of the Crick wobble hypothesis using "⇝" to denote wobbling. Here U and G can recognize not only their complements but also nucleotides which are 2-adic close (similar) to them. Note that complementary nucleotides are not 2-adic similar, because their 2-adic distance is $d_2(\cdot, \cdot) = 1$. Recognition by inosine is a special case, which can be viewed as recognition of C as complement to I and U as 2-adically similar to C. Inosine can also recognize U, although close to I but less than G. For 2-adic similarities, e.g. see Fig. 2.

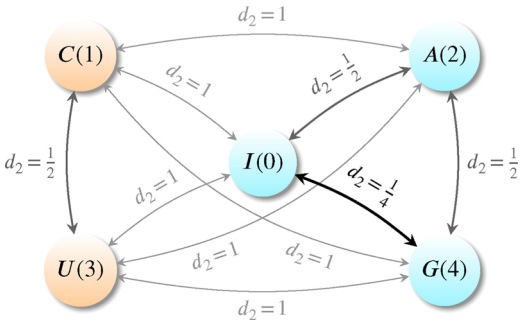

Fig. 2. This figure contains nucleotides *C*, *A*, *U*, *G*, *I* with the corresponding digits (in brackets) in their representation within codons in the 5-adic form. 2-Adic distance between all of these nucleotides is presented with the values $d_2 = 1, \frac{1}{2}, \frac{1}{4}$. According to 2-adic distance inosine and guanine are closer than inosine and adenine (note that I and G have both amino and oxo group, while adenine has only amino).

$$C \to G; \quad G \to C; \quad T \to A; \quad A \to U. \tag{17}$$

The second step in gene expression is translation from language $\mathcal{L}_2$ to language $\mathcal{L}_3$. In this case, translation of letters from $\mathcal{A}_2 \to \mathcal{A}_3$ depends on their position in words according to the next two rules.

(i) At the first two positions of nucleotides in codons, translation follows Watson-Crick base pairing

$$C \to G; \quad G \to C; \quad U \to A; \quad A \to U. \tag{18}$$

(ii) At the third position of codons translation is according to Crick's wobble hypothesis (see Fig. 1):

$$\begin{aligned} &C \to G, I; \qquad A \to U, I; \\ &U \to A, G, I; \qquad G \to C, I. \end{aligned} \tag{19}$$

From (19) one can easily see that translation at the third position is not one-to-one as in (18), but there is some ambiguity. This ambiguity enables some anticodons in tRNA to recognize more than one codon in mRNA according to wobbling over the third position of codons:

$$\begin{aligned} &C \rightsquigarrow G; \qquad A \rightsquigarrow U; \\ &U \rightsquigarrow A, G; \quad G \rightsquigarrow C, U; \quad I \rightsquigarrow C, A, U. \end{aligned} \tag{20}$$

From (20) follows that true wobble is related to *U*, *G*, and particularly to *I*.

Note that the first two nucleotides of codons in DNA are the same as the corresponding two nucleotides of anticodons in tRNA. Hence, the genetic code is a mapping from codons in DNA to amino acids.

The third step in gene expression is the translation from language $\mathcal{L}_3$ to language $\mathcal{L}_4$, that is the translation from an anticodon in tRNA to the corresponding amino acid attached to the same tRNA. Note that $\mathcal{L}_3$ has $5 \times 5 \times 5 = 125$ words but only a part is employed in translation to $\mathcal{L}_4$. This part of 125 is related to the number of tRNA capable of translating anticodons unambiguously to amino acids – in the case of the standard genetic code 31 is minimal and 41 maximal number of tRNAs. This is another example of the economic functioning of living cells. This third step is the translation of some words from $\mathcal{L}_3$ to 20 letters of $\mathcal{L}_4$. At the Table 4 is presented a reduced genetic code with minimal (31) number of tRNAs with maximal employment of inosine in tRNA anticodons.

4.2. Codon space as an ultrametric tree

In the previous Section 4.1, we introduced four languages which are sequentially used for translation of genetic information from DNA to the synthesis of proteins. The genetic code is usually considered as a translation of language $\mathcal{L}_2$ to language $\mathcal{L}_4$, i.e. translation of codons from RNA to amino acids in proteins. We want to present language $\mathcal{L}_2$, which consists of 64 codons, as an ultrametric space.

It is well known that the first two nucleotides in codons play a more important role than the third one. Almost always, when two codons have the same first two nucleotides and at the third places are purines or pyrimidines, then they code the same amino acid or stop signal. This fact indicates that 64 codons are ordered according to ultrametric distance, which is generally defined in Section 2.

Now we are introducing an example of ultrametric distance suitable to present codon space as an ultrametric tree. Namely, let the distance between any two words x and y be $d_u(x, y) = (n - m)/n$, where n is number of letters in these words, and m is the number of letters pointwise equal in both words counting from the beginning. This distance takes discrete values between 0 and 1 as follows:

$$d_u(x, y) = \frac{n-m}{n} = \begin{cases} 1, & \text{if} \quad m = 0, \\ 1 - \frac{1}{n}, & \text{if} \quad m = 1, \\ \dots & \dots \\ 0, & \text{if} \quad m = n. \end{cases} \tag{21}$$

One can easily see that smaller m implies larger distance and vice versa.

It is evident that distance (21) satisfies properties (*i*) and (*ii*) of ultrametrics (6). Property (*iii*) (i.e. strong triangle inequality) can be shown in the following way. Let distances between three words x, y and z (of the same length n) are: $d_u(x, y) = (n - m)/n$, $d_u(x, z) = (n - m_1)/n$, $d_u(y, z) = (n - m_2)/n$. Difference between these distances depends on numbers m, m_1, m_2. It is important to observe that two of these numbers m, m_1, m_2 are always equal and the third one is larger or equal to the other two. Without less of generality, let $m \geq m_1 = m_2$. Then there are the following two possibilities:

$$d_u(x, y) = d_u(x, z) = d_u(y, z) \quad \text{if} \quad m = m_1 = m_2, \tag{22}$$

$$d_u(x, y) < d_u(x, z) = d_u(y, z) \quad \text{if} \quad m > m_1 = m_2. \tag{23}$$

Properties (22) and (23) can be rewritten in the form $d_u(x, y) \leq \max\{d_u(x, z), d_u(y, z)\}$, what is just property (*iii*) in (6).

In order to apply afore introduced ultrametrics of words to codons, note that in this case $n = 3$ and $m \in \{0, 1, 2\}$ for words mutually different. Let codons x, y and z be $x = x_0x_1x_2$, $y = y_0y_1y_2$ and $z = z_0z_1z_2$, where letters are from alphabet $\mathcal{A}_1$ or $\mathcal{A}_2$. Then,

$$d_u(x, y) = \begin{cases} 1, & \text{if } x_0 \neq y_0, \\ \frac{2}{3}, & \text{if } x_0 = y_0,\ x_1 \neq y_1, \\ \frac{1}{3}, & \text{if } x_0 = y_0,\ x_1 = y_1,\ x_2 \neq y_2. \end{cases} \tag{24}$$

The corresponding strong triangle inequality can be illustrated by the following two examples:

(i) If $x_0 = y_0$ and $x_0 \neq z_0$ then
$$d_u(x, y) < d_u(x, z) = d_u(y, z). \tag{25}$$

(ii) If $x_0 = y_0$, $x_1 = y_1$ and $x_0 \neq z_0$ then
$$d_u(x, y) < d_u(x, z) = d_u(y, z). \tag{26}$$

This afore introduced ultrametricity of the codon space is illustrated by the related ultrametric tree in Fig. 3.

There is also the Baire distance, which is an example of ultrametrics and can be applied to codons to get ultrametric tree as presented in Fig. 3. For codons, it is as

$$d_B(x, y) = 2^{-m} = \begin{cases} 1, & \text{if } m = 0, \\ \frac{1}{2}, & \text{if } m = 1, \\ \frac{1}{4}, & \text{if } m = 2, \end{cases} \tag{27}$$

where m has the same meaning as in the previous example. Note that the Baire space is usually defined for infinite sequences. Instead of base 2, one can also use any other natural number different from 1.

Above two examples exhibit a rough ultrametric structuring of 64 codons according to their information meaning. Finer ultrametric

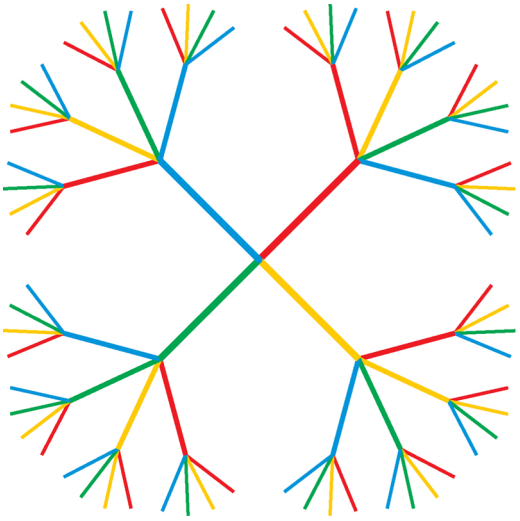

Fig. 3. This is an ultrametric tree of the genetic code. It illustrates the ultrametric structure in ordering codons with respect to their coding amino acids and stop signal in mRNA. Four colors correspond to four nucleotides with related digits: $C = 1$, $A = 2$, $U = 3$, $G = 4$. One can easily calculate ordinary ultrametric distance using formula (23) and see that distances between any three endpoints of the tree satisfy the strong triangle (ultrametric) inequality, e.g. see (25) and (26). One should look at this figure from the center to the ends of the graph.

analysis of codons can be achieved using p-adic distance between them. For example, one can adequately characterize differences between pyrimidines and purines using 2-adic distance.

4.3. p-Adic structuring of the set of codons

To apply p-adic distance to the genetic code, one has to attach adequate numbers to codons. In Section 4.1 we considered codons as three-letter words, where letters are nucleotides. This suggests construction of numbers related to codons in such a way that to digits of numbers correspond nucleotides. Acting in this direction, we have to take into account that nucleobases which are purines, as well as those which are pyrimidines, are closer than mutually a purine and a pyrimidine.

Altogether in the expression of genetic code, there are three pyrimidines (C, T, U) and three purines (A, G, I), where I denotes inosine. Keeping in mind that T (thymine) and U (uracil) are practically equivalent, then there remain five nucleotides, and we should introduce relevant five digits. In the construction of numbers for all codons, it is natural to take prime number five ($p = 5$) as a base for their representation and 0, 1, 2, 3, 4 as the corresponding digits.

Inosine is a special case and there is a sense to attach it digit 0, i.e. $I \equiv 0$. Attaching digits 1, 2, 3, 4 to rest of four nucleotides, we have to take into account that nucleobases which are purines (A, G), as well as those which are pyrimidines (C, U), have more similar meaning than a purine comparing to a pyrimidine. This similarity (closeness) naturally describes by 2-adic distance. This requirement reduces 24 possible connections between digits and nucleotides to 8, which are presented it Table 1. However, as $I \equiv 0$ one should take $A \equiv 2$, $G \equiv 4$ or $G \equiv 2$, $A \equiv 4$, because $d_2(0, 2) = d_2(4, 2) = \frac{1}{2}$. Then, it should be $C \equiv 1$, $U \equiv 3$ or $U \equiv 1$, $C \equiv 3$. By this way we have reduced $8 \to 4$ possibilities. As there is complementary base pairing $A - U$ and $C - G$, it is reasonable to take $A + U = C + G = 5$. This equality is possible in two cases: either $C \equiv 1$, $A \equiv 2$, $U \equiv 3$, $G \equiv 4$, or $U \equiv 1$, $G \equiv 2$, $C \equiv 3$, $A \equiv 4$.

Finally, we take the following identification:

$$I \equiv 0, \quad C \equiv 1, \quad A \equiv 2, \quad U(T) \equiv 3, \quad G \equiv 4. \tag{28}$$

We come to this conclusion by the following reasoning. In the wobble hypothesis, 2-adic distance between pairs is 1, i.e. maximal, except in the case of pairing inosine with adenine ($I \rightsquigarrow A$), where $d_2(I, A) = \frac{1}{2}$ what is larger than $d_2(I, G) = \frac{1}{4}$. Hence, one can say that, with respect to 2-adic distance, pairing "prefers" 2-adic dissimilarity or lesser similarity, because inosine recognizes adenine and does not pairs with guanine.

Table 1
Here are eight possible attachments between the nucleotides $\{C, A, U, G\}$ and the digits $\{1, 2, 3, 4\}$ taking care that 2-adic distance between two pyrimidines (C, U), as well as between two purines (A, G), is $\frac{1}{2}$. This is smaller than 2-adic distance between a pyrimidine and a purine, which is equal to 1. In Table 2 we used identification presented in the first row.

C = 1	A = 2	U = 3	G = 4
U = 1	G = 2	C = 3	A = 4
C = 1	G = 2	U = 3	A = 4
U = 1	A = 2	C = 3	G = 4
A = 1	C = 2	G = 3	U = 4
G = 1	U = 2	A = 3	C = 4
A = 1	U = 2	G = 3	C = 4
G = 1	C = 2	A = 3	U = 4

Now, instead of letters I, C, A, U (T), G one can use digits 0, 1, 2, 3, 4 and then languages (9)–(11) take the form:

$$\begin{aligned} &\mathcal{A}_1 = \{1, 2, 3, 4\} = \mathcal{A}_2, \\ &\mathcal{L}_1 = \{n_0 n_1 n_2 : \ n_i \in \mathcal{A}_1 = \mathcal{A}_2\}, \end{aligned} \tag{29}$$

$$\begin{aligned} &\mathcal{A}_3 = \{0, 1, 2, 3, 4\}, \\ &\mathcal{L}_3 \subset \{n_0 n_1 n_2 : \ n_i \in \mathcal{A}_3\}, \end{aligned} \tag{30}$$

where in general

$$n_0 n_1 n_2 \equiv n = n_0 + n_1 5 + n_2 5^2, \quad n_i \in \{0, 1, 2, 3, 4\}, \tag{31}$$

is a set of 125 possible words. Note that digit 0, which presents inosine, appears as $n_0 = 0$ only in anticodons of some tRNAs. When $n_i \in \{1, 2, 3, 4\}$ in $n_0 n_1 n_2$, one gets 64 standard codons. Numbers $\{0, 1, 2, 3, 4\}$ present the unique set of integers for 5-adic codon parametrization, because they are digits in expanding numbers in base 5.

Codons in Table 2 are ordered according to rising the values of their numbers presented in 5-adic expansion. As a result the corresponding amino acids are scattered on different places inside the table. Let us see what will happen if we join codons with respect to the smallest 5-adic distance. According to (8), 5-adic distance between any two different codons, $x = x_0 x_1 x_2$ and $y = y_0 y_1 y_2$, may have one of the following three values:

$$\begin{aligned} d_5(x, y) &= |(x_0 + x_1 5 + x_2 5^2) - (y_0 + y_1 5 + y_2 5^2)|_5 \\ &= \begin{cases} 1, & \text{if } x_0 \neq y_0, \\ \frac{1}{5}, & \text{if } x_0 = y_0, \ x_1 \neq y_1, \\ \frac{1}{25}, & \text{if } x_0 = y_0, \ x_1 = y_1, \ x_2 \neq y_2. \end{cases} \end{aligned} \tag{32}$$

According to the smallest 5-adic distance, which is $\frac{1}{25}$, 64 codons join into 16 quadruplets. Inside these quadruplets the first two nucleotides are the same and at the third place there are two pyrimidines (C, U) and two purines (A, G).

It is useful to use 2-adic distance inside quadruplets and it gives

$$\begin{aligned} d_2(x, y) &= |(n_0 + n_1 5 + x_2 5^2) - (n_0 + n_1 5 + y_2 5^2)|_2 \\ &= |(x_2 - y_2)25|_2 = |x_2 - y_2|_2 \\ &= \begin{cases} 1, & \text{if } x_2 - y_2 \neq \pm 2 \\ \frac{1}{2}, & \text{if } x_2 - y_2 = \pm 2. \end{cases} \end{aligned} \tag{33}$$

Note that $x_2 - y_2 = 2$ in two cases: (1) $x_2 = U = 3$, $y_2 = C = 1$ and (2) $x_2 = G = 4$, $y_2 = A = 2$. Hence, according to the smallest 2-adic distance between codons inside quadruplets, which is $\frac{1}{2}$, one obtains two doublets: one doublet has pyrimidine at the third position, while the other one has purine.

Table 2
Codons in three-letter nucleotide and three-digit 5-adic representation. In the upper part of this table there are 16 codons with inosine at the third position which plays an important role in wobble pairing. The lower part contains 64 standard codons. The ordering of codons is according to values of the corresponding numbers. Ordering according to 5-adic and 2-adic distances is presented in Table 3.

110 **CCI**	210 **ACI**	310 **UCI**	410 **GCI**
120 **CAI**	220 **AAI**	320 **UAI**	420 **GAI**
130 **CUI**	230 **AUI**	330 **UUI**	430 **GUI**
140 **CGI**	240 **AGI**	340 **UGI**	440 **GGI**
111 **CCC**	211 **ACC**	311 **UCC**	411 **GCC**
121 **CAC**	221 **AAC**	321 **UAC**	421 **GAC**
131 **CUC**	231 **AUC**	331 **UUC**	431 **GUC**
141 **CGC**	241 **AGC**	341 **UGC**	441 **GGC**
112 **CCA**	212 **ACA**	312 **UCA**	412 **GCA**
122 **CAA**	222 **AAA**	322 **UAA**	422 **GAA**
132 **CUA**	232 **AUA**	332 **UUA**	432 **GUA**
142 **CGA**	242 **AGA**	342 **UGA**	442 **GGA**
113 **CCU**	213 **ACU**	313 **UCU**	413 **GCU**
123 **CAU**	223 **AAU**	323 **UAU**	423 **GAU**
133 **CUU**	233 **AUU**	333 **UUU**	433 **GUU**
143 **CGU**	243 **AGU**	343 **UGU**	443 **GGU**
114 **CCG**	214 **ACG**	314 **UCG**	414 **GCG**
124 **CAG**	224 **AAG**	324 **UAG**	424 **GAG**
134 **CUG**	234 **AUG**	334 **UUG**	434 **GUG**
144 **CGG**	244 **AGG**	344 **UGG**	444 **GGG**

Table 3
p-Adic table of the vertebrate mitochondrial code. Concerning the smallest 5-adic distance, codons are joined into 16 quadruplets. Under the smallest 2-adic distance, these quadruplets are divided into two doublets. Each doublet codes an amino acid or stop signal. Coding is as follows: 2 amino acids (Ser, Leu) are coded by 3 doublets; 6 amino acids (Pro, Thr, Ala, Val, Arg, Gly) by 2 doublets; 12 amino acids (His, Gln, Asn, Lys, Tyr, Asp, Glu, Ile, Met, Phe, Cys, Trp) by single doublets; and stop signal (Ter) by 2 doublets. At the third position of codons, ordering of digits is 1, 3, 2, 4 to emphasize 2-adic closeness in doublets. There is the left-right symmetry around the vertical midline under interchange $1 \leftrightarrow 4$ and $2 \leftrightarrow 3$ at the first position.

CCC **111** Pro	ACC **211** Thr	UCC **311** Ser	GCC **411** Ala
CCU **113** Pro	ACU **213** Thr	UCU **313** Ser	GCU **413** Ala
CCA **112** Pro	ACA **212** Thr	UCA **312** Ser	GCA **412** Ala
CCG **114** Pro	ACG **214** Thr	UCG **314** Ser	GCG **414** Ala
CAC **121** His	AAC **221** Asn	UAC **321** Tyr	GAC **421** Asp
CAU **123** His	AAU **223** Asn	UAU **323** Tyr	GAU **423** Asp
CAA **122** Gln	AAA **222** Lys	UAA **322** Ter	GAA **422** Glu
CAG **124** Gln	AAG **224** Lys	UAG **324** Ter	GAG **424** Glu
CUC **131** Leu	AUC **231** Ile	UUC **331** Phe	GUC **431** Val
CUU **133** Leu	AUU **233** Ile	UUU **333** Phe	GUU **433** Val
CUA **132** Leu	AUA **232** Met	UUA **332** Leu	GUA **432** Val
CUG **134** Leu	AUG **234** Met	UUG **334** Leu	GUG **434** Val
CGC **141** Arg	AGC **241** Ser	UGC **341** Cys	GGC **441** Gly
CGU **143** Arg	AGU **243** Ser	UGU **343** Cys	GGU **443** Gly
CGA **142** Arg	AGA **242** Ter	UGA **342** Trp	GGA **442** Gly
CGG **144** Arg	AGG **244** Ter	UGG **344** Trp	GGG **444** Gly

4.4. *p*-Adic vertebrate mitochondrial code

In living organisms is known only a few dozens of genetic codes, despite a huge number of possibilities to connect 64 codons and 20 amino acids. In the human cells there are two codes: vertebrate mitochondrial code and standard genetic code. A mitochondrial code is in mitochondria, an organelle well-known for generation of adenosine triphosphate (ATP) which serves as the fuel for most metabolic processes. The vertebrate mitochondrial (VM) code has one of the simplest ordering of codons. Its 64 codons can be viewed as 16 quadruplets, inside of which 5-adic distance between codons is the smallest, i.e. $d_5(x, y) = \frac{1}{25}$. Each quadruplet consists of two doublets, so that 2-adic distance between codons which make doublets is $d_2(x, y) = \frac{1}{2}$. Thus, 64 codons, with respect to the successive application of 5-adic and 2-adic distance, form 32 doublets, which code 20 amino acids and stop signal. Namely, 2 amino acids are coded by three doublets, 6 amino acids and stop signal are coded by two doublets, and 12 amino acids are coded by single doublets (see Table 3). In this way, a combination of 5-adic and 2-adic distances choose doublets of codons which have the same genetic meaning.

When two doublets code the same amino acid, then they belong to the same quadruplet. When three doublets code the same amino acid (i.e. leucine or serine), then two doublets are in the same quadruplet, but the third one belongs to another quadruplet. In the case of leucine, there is the difference between the third doublet with respect to one doublet in quadruplet only in the first digit, i.e. in one doublet the first digit is 1 and in the other is 3. It follows that 2-adic distance between related codons is $d_2(132, 332) = d_2(134, 334) = \frac{1}{2}$, what means that 2-adic closeness is the same as inside quadruplet. Situation with serine is similar, and one has $d_2(241, 311) = d_2(243, 313) = \frac{1}{2}$. It is worth noting an easy way to determine divisibility of integer numbers by 2 in 5-adic expansion. Namely, if the sum of digits is even or odd, then the given number is even or odd, respectively.

4.5. *p*-Adic standard genetic code

The standard or canonical genetic code is almost universal in living organisms on our planet. It can be viewed as a slight variation of vertebrate mitochondrial code in the sense that some codons have changed amino acids assignment. Namely,

- AUA (232): Met → Ile,
- AGA (242) and AGG (244): Ter → Arg,
- UGA (342): Trp → Ter.

Thus, in the standard code there is not only coding by two, four and six codons of the same amino acid, but also by one and three codons (see Fig. 4). This is brokenness of the simple rule in the VM code: one doublet – one amino acid or stop signal. Note that brokenness of simplicity is a natural phenomenon. Namely, it is well known that the standard model of particle physics is based on spontaneously broken electroweak gauge symmetry giving mass to particles like an electron. Hence it is not unnatural that in the origin and evolution of life this kind of simplicity may become broken, giving rise to new rules appropriate to more complex and advanced organisms. Hence, one can say that the evolution of matter and life contains some evolution of initial physical and biological rules, respectively.

Comparing further the standard and VM code, one can note that still holds the following rule: if two codons code the same amino acid then they are close (similar) in respect of 2-adic or 5-adic (or both 2-adic and 5-adic) distance. This rule can be generalized to all possible variations of the genetic code as follows: two codons that code the same amino acid or stop signal are close with respect to at least one *p*-adic distance.

4.6. Euclidean representation of *p*-adic genetic code model

In previous three subsections we presented codons as some 5-adic natural numbers. It is well known that real natural numbers can be presented as points on real axis. However, there is no possibility to present *p*-adic numbers with all their properties in the Euclidean space, which geometry is based on the usual absolute value. Nevertheless, to express self-similar structure, *p*-adic numbers are often illustrated as a tree, dendrogram or fractal. These structures are scale invariant as consequence of the power-law behaviour of *p*-adic distances. Here we present 5-adic numbers,

$$x = x_0 + x_1 5 + x_2 5^2 \equiv x_0 x_1 x_2, \quad x_i \in \{0, 1, 2, 3, 4\}, \tag{34}$$

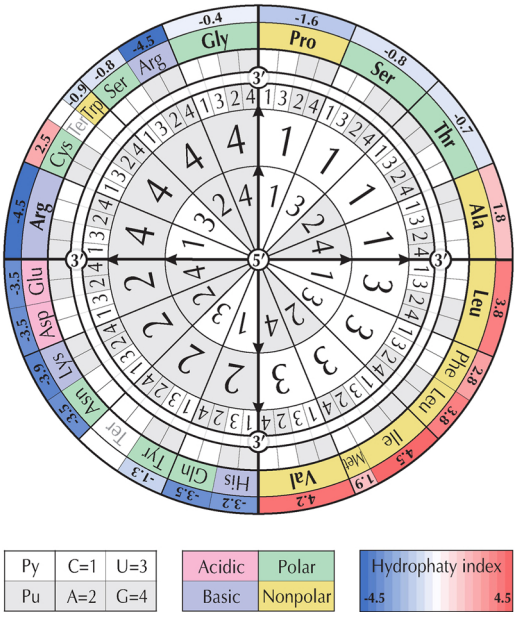

Fig. 4. Circular representation of the standard genetic code. To have pyrimidines together as well as purines, which is more compatible with 2-adic closeness, the ordering of digits is 1, 3, 2, 4 instead of 1, 2, 3, 4. Purine bases are colored gray. The hydrophobic character of the amino acid side chains is given by Kyte-Doolittle hydropathy index.

that contain codons, like fractals in plane using Euclidean distance.

Let $\mathcal{I}_5[5^3]$ be information space which points are given by (34), and let $C_5[4^3] = \{c_0c_1c_2\colon\ c_i \in \{1, 2, 3, 4\}\}$ be codon space. For p-adic amino acid space $\mathcal{A}_p$, a nontrivial p-adic representation of a codon-amino acid assignment closeness for $\mathcal{A}_5[20] \subset \mathcal{I}_5[5^3]\backslash C_5[4^3]$ 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 $a' = a_0'00 = a_0'$ and $a_0' = c_0$, where c_0 is the first base of their cognate codons.

The visual representation of p-adic information space $\mathcal{I}_5[5^3]$ and its subspaces $C_5[4^3]$ and $\mathcal{A}_5[20]$ is given in Mišić (2016), using a fractal approach based on a p-adic distance representation by usual Euclidean distance, as it is defined in Robert (2000). A formalism will be given for complete set of p-adic integers $\mathbb{Z}_p$ and then for $\mathcal{I}_5[5^3] \subset \mathbb{Z}_p$.

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

$$\varphi = \varphi_{\nu,d}\colon \mathbb{Z}_p \to F,\ \sum_{j\geq 0} x_j p^j \to \vartheta \sum_{j\geq 0} \frac{\nu(x_j)}{d^{j+1}}, \tag{35}$$

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

$$\varphi(\mathbb{Z}_p) = \bigcup_{\nu\in V}\left(\vartheta\frac{\nu}{d} + \frac{1}{d}\varphi(\mathbb{Z}_p)\right), \tag{36}$$

and thus for large enough values of d, the image $F = \varphi(\mathbb{Z}_p)$ will be a *disjoint* union of self-similar images – a fractal F (Robert, 2000).

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 self-similar, Cantorian-like set on Fig. 5. The 5-adic genetic model is represented by $\mathcal{G}_5[84] = C_5[4^3] \cup \mathcal{A}_5[20]$, while $\mathcal{I}_5[125]\backslash\mathcal{G}_5[84]$ is an unused part of information space (light gray numbers on Fig. 5).

Let $B_r(c) = \{x\colon |x - c| \leq r\}$ be a ball of radius r with center c in $\mathcal{I}_5[5^3]$. Then $\mathcal{I}_5[5^3] = B_1(c)$, where center c can be any number of $\mathcal{I}_5[5^3]$. Inside this 5-adic space $\mathcal{I}_5[5^3]$ there are 5 balls of radius $\frac{1}{5}$ and each of them contains 5 balls of radius $\frac{1}{25}$. The central ball $B_{\frac{1}{5}}(x)$ on Fig. 5 (light gray 5-adic numbers) is at maximal distance from 5-adic amino acid set and therefore does not belong to the 5-adic genetic code space.

This planar representation (Fig. 5) can be simply transformed into spatial representation by adding the third component to the digit vectors as $\nu'(0) = (0,\ 0,\ 0)$, $\nu'(1) = (-1,\ 1,\ 1)$, $\nu'(2) = (1,\ -1,\ 1)$, $\nu'(3) = (-1,\ -1,\ -1)$ and $\nu'(4) = (1,\ 1,\ -1)$. Choosing $d' = 2$ and $\vartheta' = 1$, the corresponding vector maps result in images of the Sierpinski-like tetrahedron, which is bounded by the largest tetrahedron inscribed in a unit cube centered at (0,0,0) with edges parallel to the coordinate axes. The vertical projection on the horizontal plane results in the previous planar model (Fig. 5) by omitting the third component in the digit vectors (Robert, 2000). Therefore, p-adic modeling can easily describe some of the existing genetic code models, although they were not originally defined by p-adic approach, e.g. the tetrahedral representation of the genetic code by Cristea (2002). The advantage of p-adic approach is the ability to obtain additional information on the relationship between elements using p-adic arithmetic, as in the case of finding the optimal digital mapping of genetic codes (Skutkova et al., 2019). Namely, Skutkova et al. (2019) showed that the optimal numerical map by considering all genetic codes is $C = 0$, $A = 1$, $U = 2$ and $G = 3$, what results with the same base order C, A, U, G and thus the equivalent 2-adic and 5-adic distances as in the originally proposed 5-adic model (Dragovich and Dragovich, 2006) and also here presented.

Additionally, if $\mathbf{N} = [C\ G;\ U\ A] = [1\ 4;\ 3\ 2]$ is a representative matrix of the base order with their associated 5-adic digits, then the arrangement of codon space $C_5[4^3]$ based on Euclidean representation (Fig. 5) can be readily obtained as the tensor cube $\mathbf{N}^{(3)}$ (Mišić, 2016). This means that p-adic model also inherently describes the algebraic approach based on such representative matrices, see Petoukhov (2016). All the above-mentioned features of p-adic approach make it a promising mathematical tool for analyzing some of the existing genetic code models.

4.7. *p-Adic structuring of the set of amino acids*

It is known that nucleotide in the middle position of codons is related to some chemical properties of amino acids. This means that such nucleotides play a special role in coding amino acids according to their physicochemical similarities. Such property of the genetic code can be viewed as a result of its origin and evolution. One of possibilities might be that formation of codons in tRNAs started around their middle position, see e.g. Dragovich and Dragovich (2010)

Now we want to describe this middle position nucleotide role in the framework of our p-adic formalism. To this end, one has to organize codons so that nucleotide from second position comes to the first one. A natural way to do that is to interchange places of the first and the second nucleotide in each of codons in Table 4.

Some physicochemical similarities of amino acids in Table 4 are given below.

- First strip (digit 1 in the middle position): a special case of diversity.
- Second strip (digit 2 in the middle position): average size and hydrophobic.
- Third strip (digit 3 in the middle position): average size and hydrophilic.
- Fourth strip (digit 4 in the middle position): small size and moderate in hydropathy.

It is worth noting that each anticodon in Table 4 is contained in its tRNA. Hence, p-adic distances between the numbers in Table 4 may be considered as p-adic distances between tRNAs and thus express functional similarity between tRNAs.

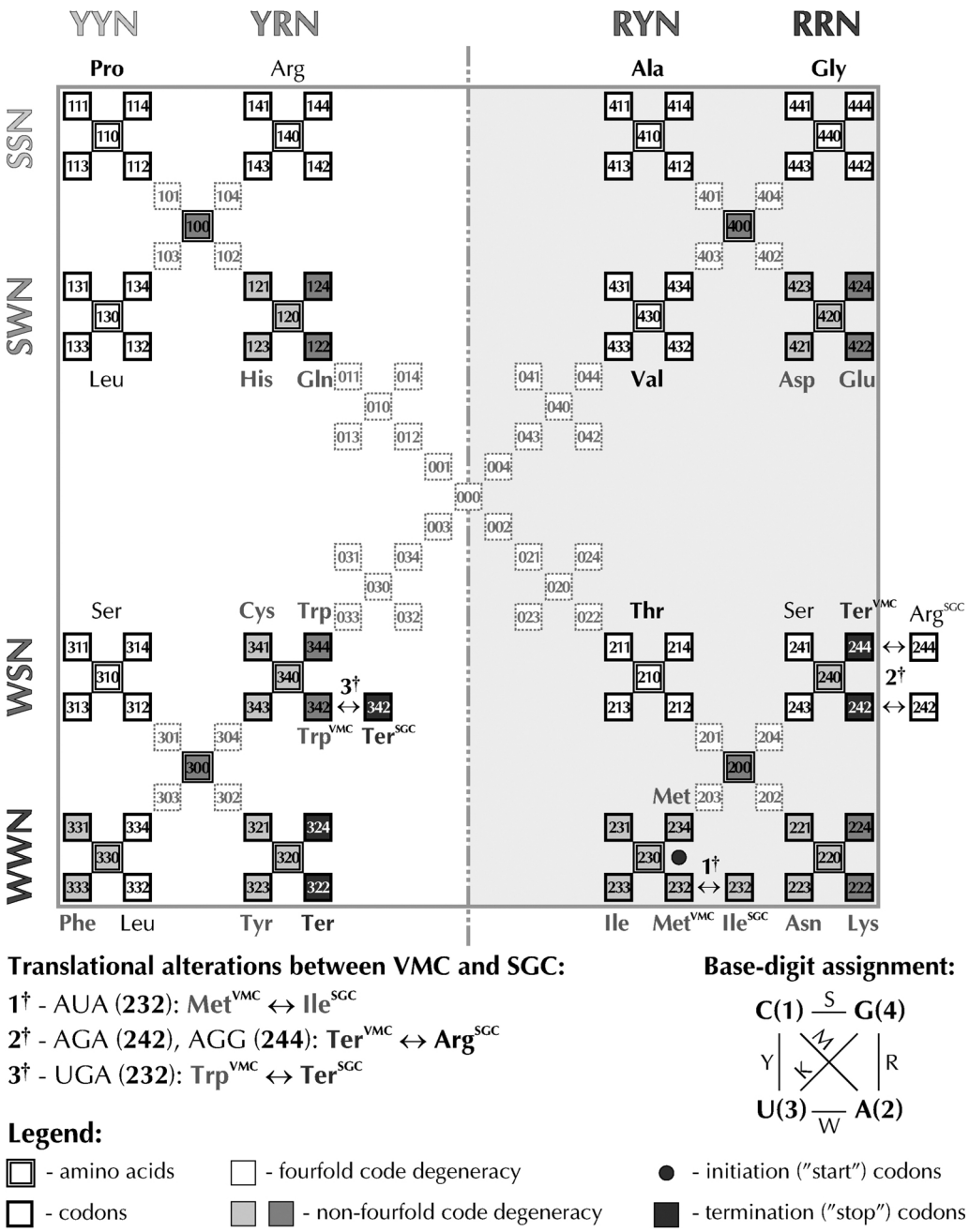

Fig. 5. Euclidean distance representation of the codon $C_5[4^3]$ and amino acid space $\mathcal{A}_5[20]$ for the *p*-adic model of the vertebrate mitochondrial (VMC) and standard (SGC) genetic code given by the translational alterations, see Mišić (2016). Abbreviations within the base-digit assignment are related to the nucleobase type: Y – pyrimidine (C,U), R – purine (A,G); S – strong (C,G), W – weak (U,A); M – amino (C,A), K – keto (U,G), while generally N stands for any nucleobase. For the fourfold code degeneracy, it is always satisfied that the 5-adic distance between the amino acid and its cognate codons is $\frac{1}{25}$ (white positions), while for non-fourfold code degeneracy there are two values of the 5-adic distance, $\frac{1}{25}$ (light gray positions) and $\frac{1}{5}$ (dark gray positions). Two halves defined by the Y/R distinction of the first codon base have an equivalent arrangement (pattern) of the fourfold and non-fourfold code degeneracy.

4.8. Genetic code as a p-adic network

In many contemporary scientific researches, networks play a very important role. Namely, many systems have the form of networks, which are the sets of nodes (vertices) joined together by links (edges). Examples mainly come from biological and social systems – from biomolecular to social networks. The above presentation gives rise to view the genetic code as a functional *p*-adic network, which can be considered as an example of the ultrametric network.

One can start from two separate systems of biomolecules – one based on 4 nucleotides and another related to 20 standard amino acids. Four types of nucleotides are chemically linked to a large number of various sequences, which are known as DNA and RNA. Standard amino acids are also chemically linked and form various peptides and proteins. By the genetic code, amino acids and stop signal are functionally linked to codons, which are the elements of a *p*-adic space.

According to Table 5, canonical amino acids can be also viewed as the elements of a *p*-adic space. Hence, one can say that the genetic code links two *p*-adic networks to one larger ultrametric network of 85 elements (64 codons + 20 amino acids + 1 stop signal). In these networks, *p*-adic distances play the role of links.

Thus the set of codons can be regarded as a network, which nodes are codons mutually linked by *p*-adic distance. Recall that there are three possibilities of 5-adic distance between codons: $\frac{1}{25}$, $\frac{1}{5}$ and 1. With respect

Table 4
An "ideal" reduced (or tRNA) standard genetic code constructed according to the Crick wobble hypothesis (the anticodons are in a 3′ to 5′ orientation, i.e. opposite to orientation in codons of mRNA). It is based on a physical connection between anticodons (language $\mathcal{L}_3$) and amino acids (alphabet $\mathcal{A}_4$) through tRNAs. Here, 31 anticodons code 20 amino acids by maximal employment of inosine (I). To each anticodon corresponds its tRNA. There are: (1) 12 amino acids (Trp, Cys, Phe, Ile, Met, Glu, Asp, Tyr, Lys, Asn, Gln, His) coded by single anticodons, (2) 5 amino acids (Gly, Val, Ala, Thr, Pro) coded by 2 anticodons, and (3) 3 amino acids (Arg, Leu, Ser) coded by 3 anticodons. It is supposed that in each strip there are four boxes with five rows (according digits 0, 1, 2, 3, 4 at the third position), where some of them do not contain anticodon with amino acid and are presented by − − −. Inside every box numbers have the same first two digits (letters), hence 5-adic distance between them is $\frac{1}{25}$. These distances between numbers can be used as distances between corresponding tRNAs to present their functional similarity.

CCI **110** Gly	- - -	- - -	GCI **410** Arg
CCC **111** Gly	ACC **211** Trp	- - -	GCC **411** Arg
- - -	- - -	- - -	- - -
- - -	- - -	UCU **313** Arg	- - -
- - -	ACG **214** Cys	UCG **314** Ser	- - -
CAI **120** Val	- - -	UAI **320** Ile	GAI **420** Leu
CAC **121** Val	- - -	UAC **321** Met	GAC **421** Leu
- - -	- - -	- - -	- - -
- - -	AAU **223** Leu	- - -	- - -
- - -	AAG **224** Phe	- - -	- - -
- - -	- - -	- - -	- - -
- - -	- - -	- - -	- - -
- - -	- - -	- - -	- - -
CUU **133** Glu	- - -	UUU **333** Lys	GUU **433** Gln
CUG **134** Asp	AUG **234** Tyr	UUG **334** Asn	GUG **434** His
CGI **140** Ala	AGI **240** Ser	UGI **340** Thr	GGI **440** Pro
CGC **141** Ala	AGC **241** Ser	UGC **341** Thr	GGC **441** Pro
- - -	- - -	- - -	- - -
- - -	- - -	- - -	- - -
- - -	- - -	- - -	- - -

to these distances, we can respectively call the corresponding subsets of codons as small, intermediate and large network community. Thus, any codon has 3 neighbors at distance $\frac{1}{25}$ and makes a small community. Any codon is also linked to 12 and 48 other codons to make an intermediate and large community, respectively. Hence, any codon belongs simultaneously to a small, an intermediate and a large community.

5. Concluding remarks

In this paper we have presented a review of various p-adic properties of the genetic code. In particular, we have emphasized ultrametric structure of the set of 64 codons with respect to their coding of amino acids and stop signal. This ultrametricity is naturally realized by subsequent action of 5-adic and 2-adic distance. Consequently, two codons which are closest under these distances code the same amino acid or stop signal. This is an exact statement in the case of the vertebrate mitochondrial code and slightly different for the standard and other genetic codes. The degeneracy of the genetic code is mainly determined by adaptors (transfer RNAs and aminoacyl-tRNA synthetases) and quite well described by p-adic distance.

Table 5
5-Adic two digit ordering of amino acids. To each of 20 amino acids two digits from Table 4 are assigned by the following way. The first digit in tRNA anticodons (0, 1, 3, 4) is removed. Then second and third positions are shifted to become the first and second positions, respectively, and are rewritten in usual order from the left to the right. Amino acids to which anticodons are removed digits 0 and 1 remain their other two digits. Amino acids which are related to the only one anticodon get a star * to the corresponding number if the removed digit is 3 or 1. This ordering follows from both the vertebrate mitochondrial and standard code. Now amino acids in the same row have some similar physicochemical properties. 5-Adic distance between two amino acids in the same row is at least $\frac{1}{5}$, while between amino acids in different rows is equal to 1.

11 Gly	12 Cys		14 Arg	12* Trp		
21 Val	22 Phe	23 Ile	24 Leu	23* Met		
31 Asp	32 Tyr	33 Asn	34 His	31* Glu	33* Lys	34* Gln
41 Ala	42 Ser	43 Thr	44 Pro			

As a new contribution, we have added our consideration of the genetic code as a sequence of translations between four corresponding languages. In the last translation, tRNAs provide a physical connection between anticodons and amino acids. In the third language we have included inosine. Consequently, one has that altogether codon and anticodon representations use: (1) six letters – 3 pyrimidines (C, T, U) and 3 purines (A, G, I), and (2) five 5-adic digits ($I = 0$, $A = 2$, $U = T = 3$, $G = 4$). Note that Cristea (2002) introduced digits for nucleotides in base 4 ($T = 0$, $C = 1$, $A = 2$, $G = 3$), however his approach is not p-adic.

Two principles. It is worth noting that in modeling of the genetic code one can use two principles: (1) principle of ultrametric codon similarity and (2) principle of nucleotide complementarity. Both of these principles have a natural realization in presented p-adic formalism.

Namely, in the vertebrate mitochondrial code, two codons which have the same first two nucleotides, and the third nucleotide in both codons is purine or pyrimidine, code the same amino acid or stop signal, and we say that such two codons are ultrametric (or p-adic) similar. Here, similarity is measured by p-adic distance: smaller distance = more close = more similar. According to this similarity, 64 codons in vertebrate mitochondria are arranged by 5-adic distance into 16 quadruplets which then split into doublets under 2-adic distance. This principle of similarity is true in the vertebrate mitochondria code and slightly broken in some other genetic codes, like the standard one.

The principle of nucleotide complementarity acts in the process of DNA replication, as well as in transcription of genes from DNA to corresponding RNA. This is realized through Watson-Crick pairing, which can be presented in terms of 5-adic digits 1, 2, 3, 4 as: $C + G = 1 + 4 = 5$ and $A + T = A + U = 2 + 3 = 5$, i.e. two nucleotides are complementary if their sum is 5. This principle is broken in the process tRNA reading of messenger RNA by Crick's wobble pairing.

Principles and their brokenness in some cases play important role in sciences. For example, some gauge symmetries are starting point in construction theory of fundamental physical interactions and brokenness of some of them give rise to generation of masses for known massive elementary particles, like electrons and protons. In the case of the genetic code, the vertebrate mitochondria act exactly by above simple principle of similarity but have remained organelles, while the vertebrates became advanced organisms although their nucleic genetic code functions with broken similarity principle. Also, brokenness of the nucleotide complementarity makes transfer of the genetic information economically. One can say that some advanced natural systems are developed as a result of brokenness of some simple principles.

Adelic approach. In this paper, in addition to 5-adic distance we also employed 2-adic one, to differ class of pyrimidines from the class of purines. For the future research, we plan to use p-adic distance with other prime numbers to express other properties of the genetic code. To do that in a systematic way, there is an appropriate mathematical tool which is related to adeles. An adele x is an infinite sequence (Gel'fand et al., 1969)

$$x = (x_\infty, x_2, x_3, \ldots, x_p, \ldots), \tag{37}$$

where x_∞ is a real number and x_p is a p-adic number with the restriction that for all but a finite set of primes p it has to be satisfied $|x|_p \leq 1$. A special case, called principal adele, is when the same integer or rational number is treated as real as p-adic for all primes. Adelic approach gives possibility to treat integers (and rational numbers) simultaneously with respect to all nontrivial norms and to get important results like formula (5). Adelic consideration should give more complete description of the genetic code.

Applications of p-adic distance to bioinformation. It is worth recalling emergence of Quantum Mechanics. At the end of 19th century there were

plenty of experimental spectroscopic data on matter radiation which produced serious difficulties to be described and understood within classical theory. Progress in description started by invention of a new physical notion – a quantum of action (equal to the Planck constant h), and later by application of new mathematical methods related to the Hilbert space. It seems that similar situation is in contemporary biology, in particular with respect to bioinformation phenomena. We believe that p-adic distance is an appropriate mathematical tool able to describe not only many properties of the genetic code but can be also successfully applied to description of many properties of other bioinformation contents.

In previous sections we have shown that similarity between codons and similarity between amino acids can be investigated by some p-adic distances. Recall that codons are ordered sequences of three nucleotides. It is natural to ask a question about application of p-adic distance in investigation of similarity (dissimilarity) between sequences longer than three nucleotides. The answer should be positive and the set of applied p-adic distances should contain all primes smaller than the largest of numbers assigned to sequences under consideration. It should be also interesting and useful to consider meaning of closeness (similarity) between such sequences regarding the relevant p-adic distances.

For example, one can consider microRNAs (miRNAs), which contain about 22 nucleotides that function by base-pairing with complementary sequence in messenger RNA. miRNAs play an important role in the regulation of gene expression. We plan to investigate p-adic distances between miRNAs and connection with their role in the gene expression regulation.

p-Adically modified the Hamming distance. Another application of p-adic distance is by modification of the Hamming one, which can be defined in the following way. Let $a = a_1\, a_2 \cdots a_n$ and $b = b_1\, b_2 \cdots b_n$ be two sequences (strings) of equal length. The Hamming distance between these two sequences is

$$d_H(a,\, b) = \sum_{i=1}^{n} d\,(a_i,\, b_i), \tag{38}$$

where $d(a_i, b_i) = 0$ if $a_i = b_i$, and $d(a_i, b_i) = 1$ if $a_i \neq b_i$. In other words, $d_H(a, b) = n - \nu$, where ν is the number of positions at which elements of both sequences are equal.

We introduce a p-adic Hamming distance as

$$d_{\mathrm{pH}}(a,\, b) = \sum_{i=1}^{n} d_p(a_i,\, b_i), \tag{39}$$

where $d_p(a_i, b_i) = |a_i - b_i|_p$ is p-adic distance between numbers a_i and b_i. When $a_i,\, b_i \in \mathbb{N}$ then $d_p(a_i, b_i) \leq 1$. If also $a_i - b_i \neq 0$ is divisible by p then $d_p(a_i, b_i) < 1$. The following inequality holds: $d_{pH}(a, b) \leq d_H(a, b)$. Note that constructing p-adic modified Hamming distance the corresponding elements of sequences a and b have to be some numbers, while in the ordinary Hamming case they can be of arbitrary nature.

In the case of sequences as parts of DNA and RNA, this modified distance is finer and should be more appropriate than the ordinary Hamming one. Elements in (39) can be nucleotides or codons.

p-Adic approach should be also useful for investigation of similarity between some other sequences which do not belong to DNA, RNA or proteins. These sequences can be related to other systems of molecules in the cells which are carriers of some bioinformation.

We also plan to employ this p-adic approach to creation of an artificial language (Dragovich, 2019) and investigation of similarity between words in some human languages. It seems interesting further investigation of p-adic approach to PAM matrix (Khrennikov and Kozyrev, 2009) and evolution of the genetic code, see Dragovich and Dragovich (2010) and Avetisov and Zhuravlev (2007). Application to cognitive neuroscience is a big challenge (Iurato et al., 2016).

Conflict of interest

All authors declare that they have no conflict of interest in this research.

Acknowledgements

The authors are thankful to M. Rakočević for useful discussions on many aspects of the genetic code. They also thank reviewers for comments towards better presentation of this paper. B. D. thanks M. Barbieri for inspiring communications on Code Biology. This paper is a result of the research funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, projects: OI 173052, OI 174012, TR 32040 and TR 35023.

References

Avetisov, V.A., Zhuravlev, Y.N., 2007. An evolutionary interpretation of the p-adic ultrametric diffusion equation. Dokl. Math. 75, 453–455. https://doi.org/10.1134/S1064562407030325.
Barbieri, M., 2015. Evolution of the genetic code: the ribosome-oriented model. Biol. Theory 10, 301–310. https://doi.org/10.1007/s13752-015-0225-z.
Barbieri, M., 2018. What is code biology? Biosystems 164, 1–10. https://doi.org/10.1016/j.biosystems.2017.10.005.
Barbieri, M., 2019. Evolution of the genetic code: the ambiguity-reduction theory. BioSystems (this issue).
Bashford, J., Tsohantjis, I., Jarvis, P., 1997. Codon and nucleotide assignments in a supersymmetric model of the genetic code. Phys. Lett. A 233, 481–488. https://doi.org/10.1016/S0375-9601(97)00475-1.
Bernfield, M.R., Nirenberg, M.W., 1965. RNA codewords and protein synthesis. Science 147, 479–484. https://doi.org/10.1126/science.147.3657.479.
Brekke, L., Freund, P.G.O., 1993. p-adic numbers in physics. Phys. Rep. 233, 1–66. https://doi.org/10.1016/0370-1573(93)90043-D.
Crick, F.H., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. U.S.A. 43, 416–421. https://doi.org/10.1073/pnas.43.5.416.
Crick, F.H.C., 1966. Codon-anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19, 548–555. https://doi.org/10.1016/S0022-2836(66)80022-0.
Crick, F.H.C., 1968. The origin of the genetic code. J. Mol. Biol. 38, 367–379. https://doi.org/10.1016/0022-2836(68)90392-6.
Cristea, P.D., 2002. Conversion of nucleotides sequences into genomic signals. J. Cell. Mol. Med. 6, 279–303. https://doi.org/10.1111/j.1582-4934.2002.tb00196.x.
Downes, A.M., Richardson, B.J., 2002. Relationships between genomic base content and distribution of mass in coded proteins. J. Mol. Evol. 55, 476–490. https://doi.org/10.1007/s00239-002-2343-z.
Dragovich, B., 2012. p-Adic structure of the genetic code. arXiv:1202.2353.
Dragovich, B., 2016. Genetic code and number theory. FU Phys. Chem. Tech. 14, 225–241. https://doi.org/10.2298/FUPCT1603225D. arXiv:0911.4014v1.
Dragovich, B., 2019. Towards artificial p-adic language. Filomat 33 (4) (in press).
Dragovich, B., Dragovich, A., 2006. A p-adic model of DNA sequence and genetic code. arXiv:q-bio/0607018v1.
Dragovich, B., Dragovich, A., 2007. p-Adic degeneracy of the genetic code. In: Dragovich, B., Rakić, Z. (Eds.), Proc. Conf.: Modern Mathematcal Physics. Institute of Physics, Belgrade, pp. 179–188. arXiv:0707.0764v1.
Dragovich, B., Dragovich, A., 2010. p-Adic modelling of the genome and the genetic code. Comput. J. 53, 432–442. https://doi.org/10.1093/comjnl/bxm083. arXiv:0707.3043v1.
Dragovich, B., Dragovich, A.Y., 2009. A p-adic model of DNA sequence and genetic code. p-Adic Numbers Ultrametr. Anal. Appl. 1, 34–41. https://doi.org/10.1134/S2070046609010038. arXiv:0607018v1.
Dragovich, B., Khrennikov, A.Y., Kozyrev, S.V., Volovich, I.V., 2009. On p-adic mathematical physics. p-Adic Numbers Ultrametr. Anal. Appl. 1, 1–17. https://doi.org/10.1134/S2070046609010014. arXiv:0904.4205v1.
Dragovich, B., Khrennikov, A.Y., Kozyrev, S.V., Volovich, I.V., Zelenov, E.I., 2017a. p-Adic mathematical physics: the first 30 years. p-Adic Numbers Ultrametr. Anal. Appl. 9, 87–121. https://doi.org/10.1134/S2070046617020017. arXiv:1705.04758.
Dragovich, B., Khrennikov, A.Y., Mišić, N.Ž., 2017b. Ultrametrics in the genetic code and the genome. Appl. Math. Comput. 309, 350–358. https://doi.org/10.1016/j.amc.2017.04.012. arXiv:1704.04194v1.
Dragovich, B., Mišić, N.Ž., 2018. p-Adic side of the genetic code and the genome. In: Mondaini, R.P. (Ed.), Trends in Biomathematics: Modeling, Optimization and Computational Problems. Springer International Publishing., pp. 75–89. https://doi.org/10.1007/978-3-319-91092-5_6.
Fimmel, E., Strüngmann, L., 2018. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems 164, 186–198. https://doi.org/10.1016/j.biosystems.2017.09.007.
Fimmel, E., Strüngmann, L., 2016. Yury Borisovich Rumer and his 'biological papers' on the genetic code. Philos. Trans. Royal Soc. A 374, 20150228. https://doi.org/10.1098/rsta.2015.0228.
Finkelstein, A.V., Ptitsyn, O.B., 2002. Protein Physics. Soft Condensed Matter, Complex Fluids and Biomaterials. Academic Press, London. https://doi.org/10.1016/B978-0-12-256781-0.X5000-8.
Forger, M., Sachse, S., 2000. Lie superalgebras and the multiplet structure of the genetic code. I. Codon representations. J. Math. Phys. 41, 5407–5422. https://doi.org/10.1063/1.533417. arXiv:9808001.
Frappat, L., Sciarrino, A., Sorba, P., 2001. Crystallizing the genetic code. J. Biol. Phys. 27, 1–38. https://doi.org/10.1023/A:1011874407742. arXiv:0003037v1.
Gamow, G., 1954. Possible relation between deoxyribonucleic acid and protein

structures. Nature 173, 318. https://doi.org/10.1038/173318a0.
Gel'fand, I.M., Graev, M.I., Pyatetskii-Shapiro, I.I., 1969. Representation Theory and Automorphic Functions. Saunders Mathematics Books, Saunders, Philadelphia.
Hayes, B., 1998. Computing science: the invention of the genetic code. Am. Sci. 86, 8–14.
Hornos, J.E.M., Hornos, Y.M.M., 1993. Algebraic model for the evolution of the genetic code. Phys. Rev. Lett. 71, 4401–4404. https://doi.org/10.1103/PhysRevLett.71.4401.
Iurato, G., Khrennikov, A., Murtagh, F., 2016. Formal foundations for the origins of human consciousness. *p*-Adic Numbers Ultrametr. Anal. Appl. 8, 249–279. https://doi.org/10.1134/S2070046616040014.
Khrennikov, A., Kozyrev, S., 2007. Genetic code on the diadic plane. Physica A: Stat. Mech. Appl. 381, 265–272. https://doi.org/10.1016/j.physa.2007.03.018. arXiv:0701007v3.
Khrennikov, A., Kozyrev, S., 2009. 2-Adic clustering of the PAM matrix. J. Theor. Biol. 261, 396–406. https://doi.org/10.1016/j.jtbi.2009.08.014. arXiv:0903.0137v3.
Koonin, E.V., Novozhilov, A.S., 2009. Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61, 99–111. https://doi.org/10.1002/iub.146.
Kun, Á., Radványi, Á., 2018. The evolution of the genetic code: impasses and challenges. Biosystems 164, 217–225. https://doi.org/10.1016/j.biosystems.2017.10.006.
Mišić, N.Ž., 2014. From genetic code toward spacetime geometry. In: Dragovich, B., Panajotović, R.T.D. (Eds.), Proc. TABIS. 2013 Conf.: Theoretical Approaches to BioInformation Systems. Institute of Physics, Belgrade, pp. 101–123. http://www.tabis2013.ipb.ac.rs/tabis2013.pdf.
Mišić, N.Ž., 2016. Standard genetic code: *p*-Adic modelling, nucleon balances and self-similarity. FU Phys. Chem. Tech. 14, 275–298. https://doi.org/10.2298/FUPCT1603275M.
Petoukhov, S.V., 2016. The system-resonance approach in modeling genetic structures. Biosystems 139, 1–11. https://doi.org/10.1016/j.biosystems.2015.11.001.
Rakocevic, M.M., 2004. A harmonic structure of the genetic code. J. Theor. Biol. 229, 221–234. https://doi.org/10.1016/j.jtbi.2004.03.017.
Rammal, R., Toulouse, G., Virasoro, M.A., 1986. Ultrametricity for physicists. Rev. Mod. Phys. 58, 765–788. https://doi.org/10.1103/RevModPhys.58.765.
Robert, A.M., 2000. A Course in *p*-adic Analysis. Graduate Texts in Mathematics. Springer, New York.
Rumer, Y.B., 1966. Systematization of codons in the genetic code. Dokl. Akad. Nauk SSSR 167, 1393–1394 (in Russian).
Rumer, Y.B., 1968. Systematization of codons in the genetic code. Dokl. Akad. Nauk SSSR 183, 225–226 (in Russian).
Schikhof, W.H., 1985. Ultrametric Calculus: An Introduction to *p*-Adic Analysis. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge. https://doi.org/10.1142/1581.
Shcherbak, V., 2008. The Arithmetical Origin of the Genetic Code. Springer, Dordrecht, pp. 153–185. https://doi.org/10.1007/978-1-4020-6340-4_7.
Shcherbak, V.I., 1993. Twenty canonical amino acids of the genetic code: the arithmetical regularities. Part I. J. Theor. Biol. 162, 399–401. https://doi.org/10.1006/jtbi.1993.1096.
Shcherbak, V.I., 1994. Sixty-four triplets and 20 canonical amino acids of the genetic code: the arithmetical regularities. Part II. J. Theor. Biol. 166, 475–477. https://doi.org/10.1006/jtbi.1994.1042.
Shcherbak, V.I., 2003. Arithmetic inside the universal genetic code. Biosystems 70, 187–209. https://doi.org/10.1016/S0303-2647(03)00066-2.
Skutkova, H., Maderankova, D., Sedlar, K., Jugas, R., Vitek, M., 2019. A degeneration-reducing criterion for optimal digital mapping of genetic codes. Comput. Struct. Biotechnol. J. 17, 406–414. https://doi.org/10.1016/j.csbj.2019.03.007.
Swanson, R., 1984. A unifying concept for the amino acid code. Bull. Math. Biol. 46, 187–203. https://doi.org/10.1016/S0092-8240(84)80018-X.
Vladimirov, V.S., Volovich, I.V., Zelenov, E.I., 1994. *p*-Adic Analysis and Mathematical Physics. World Scientific, Singapore. https://doi.org/10.1142/1581.
Watson, J.D., Crick, F.H.C., 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171, 737–738. https://doi.org/10.1038/171737a0.
Woese, C.R., 1965. Order in the genetic code. Proc. Natl. Acad. Sci. U.S.A. 54, 71–75. https://doi.org/10.1073/pnas.54.1.71.
Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. 74, 5088–5090.