Filomat 33:4 (2019), 1227–1233 https://doi.org/10.2298/FIL1904227D



Published by Faculty of Sciences and Mathematics, University of Niš, Serbia Available at: http://www.pmf.ni.ac.rs/filomat

Towards p-Adic Artificial Language

Branko Dragovich^{a,b}

^aInstitute of Physics, University of Belgrade,Pregrevica 118, 11080 Belgrade, Serbia ^bMathematical Institute of Serbian Academy of Sciences and Arts, Kneza Mihaila 36, Belgrade, Serbia

Abstract. Motivated by successful *p*-adic modeling of the genetic code, in this paper a *p*-adic (ultrametric) language is introduced. This language can serve as an effective and advanced constructive element of a future artificial intelligence. In this geometric approach to artificial language, we mainly use *p*-adic distance as the most powerful example of ultrametrics. Here *p*-adic distance is a tool to quantify similarity between words – similarity between their structure and similarity between their meaning. As a simple illustration, a *p*-adic language with four letters and their 3-letter words is presented.

1. Introduction

Language, as finite ordered sequences (words) of some finite set of elements (letters) with relevant rules, is an important component of all living organisms – from single cells to the humans. There are different forms of languages that have emerged during evolution.

The oldest language is related to the genetic code. Namely, the genetic code can be viewed as a language, which letters are four nucleotides (C, A, U, G) and words are codons as ordered triples of nucleotides. Meaning of 64 codons are 20 canonical amino acids plus one stop signal. Consequently, meaning of genes are proteins. This genetic language emerged at the very beginning of life on our planet and exists so far in almost the same form without changes. It is an invariant of life evolution, because all other constituents of living organisms have experienced changes.

In a series of papers [1-6] (see also [7]) was shown that set of 64 codons can be presented as an ultrametric tree and that the degeneracy of the vertebrate mitochondrial (VM) code can be adequately described by *p*-adic distance. The VM code is very simple and other dozen genetic codes can be considered as its slight variations.

Note that *p*-adic distance is the most elaborated and useful example of ultrametrics. By appropriate identification of nucleotides and digits in 5-adic representation of 64 codons, one obtains 32 codon doublets which code amino acids or stop signal. Codons in these doublets are at the smallest 5-adic and 2-adic distance. One can say that two codons coding the same amino acid or stop signal have the same meaning, because they are *p*-adically (ultrametrically) closest each other with respect to all other codons. For such codons we also say that they are, in the information sense, similar.

²⁰¹⁰ Mathematics Subject Classification. Primary 11F85, 68T99

Keywords. p-adic distance, p-adic similarity, p-adic artificial language

Received: 04 Oktober 2018; Accepted: 19 December 2018

Communicated by Ljubica S. Velimirović

This paper was supported in part by Ministry of Education, Science and Technological Development of the Republic of Serbia, grant No 174012.

Email address: dragovich@ipb.ac.rs (Branko Dragovich)

The *p*-adic (ultrametric) approach to closeness between codons was extended to canonical amino acids, and also between codons and amino acids, see e.g. [5]. In this way, the genetic code can be regarded as an ultrametric network.

It is worth noting that doublets of vertebrate mitochondrial codons, which code the same amino acid or stop signal, have the same first two nucleotides. In fact, two sequences of letters are more close if they have more same elements at places counting from the beginning. This is a general property of any ultrametric space of sequences. In human languages, which are based on letters, there are also some ultrametricity, but not so strict as in the case of codons.

In this paper, we propose an ultrametric (*p*-adic) language for a future artificial intelligence. In biology there is an approximative rule between physico-chemical structure and biological function: similar in structure – similar in function. Our goal is to have an artificial language with strong connection between structure and meaning: *similar in structure – similar in meaning*. Such language with ultrametric similarity, should be simple, economic and effective. Namely, our intention is to have language in which notions with similar meaning are expressed by words of similar structure. Therefore, we consider an ultrametric approach which is motivated by language of the genetic code.

2. Ultrametric (p-Adic) Space of Sequences

Ultrametric space is a set U with a real-valued function *d* of any two elements on U satisfying the following properties:

$$\begin{array}{l} (i) \ d(x,y) \geq 0, \ d(x,y) = 0 \ \text{iff} \ x = y, \\ (ii) \ d(x,y) = d(y,x), \\ (iii) \ d(x,y) \leq \max\{d(x,z), d(z,y)\}. \end{array}$$

$$(1)$$

Property (iii) is called *strong triangle inequality*. Note that usual distance must satisfy triangle inequality, i.e.

$$d(x, y) \le d(x, z) + d(z, y).$$
 (2)

As a consequence of property (iii), all ultrametric spaces have some unusual properties, like:

- in all triangles two sides are equal and the third one is smaller or equal to others,
- there is no partial intersection of the balls,
- any point of a ball can be its center.

There are many examples of ultrametric spaces. In the sequel we shall consider ultrametric space of sequences. Let S be a set of sequences of the same finite length *n*. The set S can be a finite ultrametric space by defining an appropriate distance.

An *ordinary ultrametric distance* between any two elements *x* and *y* of S can be introduced in the following way: d(x, y) = n - m, where *m* is a number of the first equal elements counting from the beginning in both sequences *x* and *y*. Thus $0 \le m \le n$ and $0 \le d(x, y) \le n$. It is evident that properties (*i*) and (*ii*) are satisfied. Property (*iii*) can be shown by analyzing possible distances between three arbitrary sequences *x*, *y* and *z*. It is worth noting that this ordinary ultrametric distance can be redefined by scaling it as $d_s(x, y) = \frac{n-m}{n}$ and then it takes values $0 \le d_s(x, y) \le 1$.

Another example of ultrametric space of sequences is related to the *Baire distance*. This distance is defined as $d_B(x, y) = 2^{-m}$, where *m* is the same as in the previous case, i.e. it is a number of the first positions at which sequences *x* and *y* have the same elements. Here the base 2 can be replaced by any other natural number greater than 2. In the definition of the Baire distance is not necessary that sequences are finite and equal in length.

2.1. p-Adic distance

The most significant example of ultrametrics is p-adic distance. Consider now finite sequences of the fixed length n in the form

$$x = x_0 x_1 x_2 \dots x_n \,, \tag{3}$$

where $x_i = 0, 1, ..., p - 1$ (are integers between 0 and p - 1) and p is a prime number. Then these sequences can be viewed as numbers in the base p in the following way:

$$x = x_0 + x_1 p + x_2 p^2 + \dots + x_n p_n \equiv x_0 x_1 x_2 \dots x_n ,$$
(4)

where x_i are related digits. By definition, *p*-adic distance between $x = x_0 x_1 x_2 \dots x_n$ and $y = y_0 y_1 y_2 \dots y_n$ is

$$d_p(x,y) = p^{-m}, (5)$$

where *m* is now number of equal first digits in both *x* and *y*. In other words, if difference between two integers *x* and *y* is divisible by p^m , then *p*-adic distance between them is p^{-m} . Thus, one can say: *more divisibility* = *smaller distance* = *more similarity*. *p*-Adic distance can be applied to some ultrametric sequences only when it is possible to transform them to numbers in a base *p*, where *p* is a prime number. Numbers endowed with *p*-adic distance are called *p*-adic numbers. In the last 30 years *p*-adic numbers, and related *p*-adic analysis, have had many applications in construction of new models in physics and biology (for a recent review, see [8]). In particular, *p*-adic models play are appropriate for description of complex systems with hierarchical structure.

2.2. On real and p-adic numbers

It is worth noting some differences between real and p-adic numbers. To see it in a simple way, it is sufficient to consider only natural numbers. Any natural number n has a unique expansion with respect to a given prime number p, i.e.

$$n = n_0 + n_1 p + n_2 p^2 + \dots + n^k p^k, (6)$$

where $n_i = 0, 1, 2, ..., p - 1$ (i = 0, 1, 2, ..., k) are digits. There are two kinds of absolute value (norm) for this natural number:

- ordinary absolute value, usually denoted by $|\cdot|$,
- *p*-adic absolute value (*p*-adic norm), denoted by $|\cdot|_p$.

Ordinary absolute value of this number is just usual sum of its terms, i.e. |n| = n. However, in the case of *p*-adic absolute value, situation is quite different and in a sense inverse. Namely, by definition, *p*-adic absolute value of *n*, presented as $n = m p^k$ with respect to a given *p* and *m* is not divisible by *p*, is

$$|n|_{p} = |mp^{k}| = p^{-k}.$$
(7)

Hence, *p*-adic absolute value of *n* in (6) is equal to the highest *p*-adic value of its terms. *p*-Adic absolute value is a very important example of non-Archimedean (ultrametric) norm and satisfies the following properties:

1)
$$|x|_p \ge 0$$
, $(|0|_p = 0)$, 2) $|xy|_p = |x|_p |y|_p$, 3) $|x + y|_p \le \max\{|x|_p, |y|_p\}$, (8)

where here *x*, *y*, *z* are some natural (rational) numbers. This *p*-adic norm can be easily extended to integer and also rational numbers. From algebraic point of view, set of rational numbers \mathbb{Q} is a simple infinite number field. All numerical results of measurements are some rational numbers. According to the Ostrowski theorem, any nontrivial norm on the field \mathbb{Q} is equivalent either to ordinary absolute value or to *p*-adic absolute value, where *p* is a prime number. As the field \mathbb{R} of real numbers can be obtained by completion of \mathbb{Q} with respect to distance induced by the ordinary absolute value, so the *p*-adic number field \mathbb{Q}_p is

completion by the same procedure where instead of ordinary absolute value is used *p*-adic norm, see e.g. [9]. There is *p*-adic analysis, which has been employed in construction of some physical models since 1987 and is named *p*-adic mathematical physics (for a recent review, see [8]). *p*-Adic modeling in physics was extended to bioinformation, and *p*-adic model of the genetic code [1–6] is an evident successful example.

Using positional notation, we present number (6) differently in real and *p*-adic case: $n = n_k n_{k-1} \dots n_1 n_0 -$ real number, and $n = n_0 n_1 \dots n_{k-1} n_k - p$ -adic number. By this presentation, in both cases we see that values of terms decrease from the left to the right of digit sequences.

Real and *p*-adic distance are related to their absolute values respectively as follows: d(x, y) = |x - y|, $d_p(x, y) = |x - y|_p$, where *x* and *y* are some natural (rational) numbers.

3. Ultrametric (p-Adic) Language

We want to construct now an ultrametric language in the following way. Let we have *m* letters of an alphabet and let *k* be a fixed length of the corresponding words. Then $n = m^k$ is a total number of words in this language. Let these *n* words denote some *n* things (objects). Can we, in principle, assign a number of properties to each of these things (words) and to be able to describe them? The answer is positive and is related to the *p*-adic absolute value. As it is well known, there are infinitely many prime numbers. To a natural number we can apply any *p*-adic norm, which is less or equal to 1. For a fixed *p*, let its *p*-adic norm is related to a definite characteristic, which we can call *p*-adic property. For different primes we will have different *p*-adic properties. Using *p*-adic distance between words we can investigate similarity of *p*-adic properties between these words. To illustrate this subject, let us consider an example where number of letters *m* = 4 and length of words *k* = 3. There are 64 words for 64 objects: A_1 , A_2 , A_3 , ..., A_{63} , A_{64} . Let these 64 words (objects) be presented in base 5 as follows:

$$x = x_0 + x_1 5 + x_2 5^2 \equiv x_0 x_1 x_2, \tag{9}$$

where $x_i = 1, 2, 3, 4$. Note that 5-adic norm of any 64 numbers of the form x in (9) is $|x|_5 = 1$, because the first digit $x_0 \neq 0$. But 5-adic distance between two different numbers $a = a_0 + a_1 5 + a_2 5^2$ and $b = b_0 + b_1 5 + b_2 5^2$ may have one of the following three values:

$$d_5(a,b) = \begin{cases} 1 & \text{if } a_0 \neq b_0, \\ 1/5 & \text{if } a_0 = b_0 \text{ and } a_1 \neq b_1, \\ 1/25 & \text{if } a_0 = b_0, a_1 = b_1 \text{ and } a_2 \neq b_2. \end{cases}$$
(10)

At Table 1 we presented 64 numbers in base 5 as a numerical representation of 64 words, which denote some objects A_1 , A_2 , A_3 , ..., A_{63} , A_{64} . With respect to 5-adic distance, for each object A_i there are 48 other objects at distance equal 1, also 12 objects at distance $\frac{1}{5}$, and 3 objects at distance $\frac{1}{25}$. For example, object A_1 presented by number 111 is at 5-adic distance equal 1 with respect to all other numbers which first digit is 2, 3 or 4.

Note that in *p*-adic model of the genetic code (vertebrate mitochondrial code), Table 1 is used for modeling 5-adic and 2-adic similarity between codons to find rule according to which two codons code the same amino acid (for details see, e.g. [5]).

At Table 2 all 64 numbers, from Table 1, are presented as product of their primes written in the decimal notation. Thus, for each number we use here two bases: on the left hand side of equality is base 5 and on the right hand side notation of primes is in base 10. One can see presence of all primes between 2 and 61 except 5, as a consequence of given representation in base 5. This presentation by product of related primes gives us possibility to find easily *p*-adic norms according to definition (7) for all 64 natural numbers. When there is not explicitly a prime number in the factorization of presented numbers, it can be regarded that such prime number is present like $p^0 = 1$ and the corresponding *p*-adic norm is 1. So all 64 numbers with respect to primes p > 61 have *p*-adic norm equal to 1.

It is also important to know *p*-adic distances between any two numbers in Table 1. As a consequence of representation in base 5 with digits 1, 2, 3, 4 it is rather easy to find some distances for the cases p = 2 and p = 3. For example $d_2(A_i + 16, A_i) = 1$, $d_2(A_i + 32, A_i) = 2^{-1}$ and $d_3(A_i + 48, A_i) = 3^{-1}$.

111 /	211 4	211 /	111 A
$ III A_1 $	211 A ₂	511 A3	411 A4
112 A ₁₇	212 A_{18}	312 A_{19}	412 A_{20}
113 A ₃₃	213 A ₃₄	313 A ₃₅	413 A ₃₆
114 A ₄₉	214 A_{50}	$314 A_{51}$	414 A ₅₂
121 A_5	221 A_6	321 A ₇	421 A ₈
122 A ₂₁	222 A ₂₂	322 A ₂₃	422 A ₂₄
123 A ₃₇	223 A ₃₈	323 A ₃₉	423 A ₄₀
124 A ₅₃	224 A_{54}	$324 A_{55}$	424 A_{56}
131 A ₉	231 A_{10}	331 A_{11}	431 A ₁₂
132 A ₂₅	232 A_{26}	332 A ₂₇	432 A ₂₈
133 A ₄₁	233 A ₄₂	333 A ₄₃	433 A ₄₄
134 A ₅₇	234 A_{58}	334 A ₅₉	434 A ₆₀
141 A ₁₃	241 A ₁₄	$341 A_{15}$	441 A ₁₆
142 A ₂₉	242 A_{30}	342 A ₃₁	442 A ₃₂
143 A ₄₅	243 A_{46}	343 A ₄₇	443 A ₄₈
144 A ₆₁	244 A_{62}	$344 A_{63}$	444 A_{64}

Table 1: This table contains 64 numbers presented by ordered digits 1, 2, 3, 4 of length 3 in base 5. The corresponding words, which denote objects, are denoted by *A* with indices increasing with real values of the related numbers. It is worth noting that 5-adic distance between numbers in every row is 1, and that in every column there are four quadruplets which numbers are at 5-adic distance $\frac{1}{25}$. For example, if 5-adic property means color then there are 64 colored objects, which can be classified into three classes: very similar ($d_5 = \frac{1}{25}$, in quadruplets where the first two digits are the same), a little similar ($d_5 = \frac{1}{5}$, when the first digits are the same and second are different) and not similar ($d_5 = 1$, when the first digits are different).

111 = 31	$211 = 2^5$	$311 = 3 \times 11$	$411 = 2 \times 17$
$112 = 2^3 \times 7$	$212 = 3 \times 19$	$312 = 2 \times 29$	412 = 59
$113 = 3^4$	$213 = 2 \times 41$	313 = 83	$413 = 2^2 \times 3 \times 7$
$114 = 2 \times 53$	214 = 107	$314 = 2^2 \times 3^3$	414 = 109
$121 = 2^2 \times 3^2$	221 = 37	$321 = 2 \times 19$	$421 = 3 \times 13$
122 = 61	$222 = 2 \times 31$	$322 = 3^2 \times 7$	$422 = 2^6$
$123 = 2 \times 43$	$223 = 3 \times 29$	$323 = 2^3 \times 11$	423 = 89
$124 = 3 \times 37$	$224 = 2^4 \times 7$	324 = 113	$424 = 2 \times 3 \times 19$
131 = 41	$231 = 2 \times 3 \times 7$	331 = 43	$431 = 2^2 \times 11$
$132 = 2 \times 3 \times 11$	232 = 67	$332 = 2^2 \times 17$	$432 = 3 \times 23$
$133 = 7 \times 13$	$233 = 4 \times 23$	$333 = 3 \times 31$	$433 = 2 \times 47$
$134 = 2^2 \times 29$	$234 = 3 \times 39$	$334 = 2 \times 59$	$434 = 7 \times 17$
$141 = 2 \times 23$	241 = 47	$341 = 2^4 \times 3$	$441 = 7^2$
142 = 71	$242 = 2^3 \times 3^2$	342 = 73	$442 = 2 \times 37$
$143 = 2^5 \times 3$	243 = 97	$343 = 2 \times 7^2$	$443 = 3^2 \times 11$
$144 = 11^2$	$244 = 2 \times 61$	$344 = 3 \times 41$	$444 = 2^2 \times 31$

Table 2: Every number in this table is presented by two ways and connected by the sign of equality "=". On the left hand side of the equality is a number in base 5 presented by digits 1,2,3,4. On the right hand side is the same number presented by product of its primes in decimal notation. These products of primes help us to easily find related *p*-adic norms. Note that all primes, except 5, between 2 and 61 are employed in this table.

1232

3.1. Adelic aspects

Note that any of the above 64 numbers, as well as any rational number, has some real and *p*-adic properties, for all infinitely many primes. In mathematics there is way to treat these real and *p*-adic properties simultaneously. It is related to the notion of adeles and ideles. An *adele* α is an infinite sequence

$$\alpha = (\alpha_{\infty}, \alpha_2, \alpha_3, \cdots, \alpha_p, \cdots), \quad \alpha_{\infty} \in \mathbb{R}, \ \alpha_p \in \mathbb{Q}_p,$$
(11)

where for all but a finite set \mathcal{P} of primes p one has that $\alpha_p \in \mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p \le 1\}$, which is the ring of p-adic integers.

Here, it is sufficient to know what is a principal idele. Principal idele ρ_{id} is an infinite sequence

$$\rho_{id} = (\rho_{\infty}, \rho_2, \rho_3, \rho_5, ..., \rho_p, ...),$$
(12)

where all ρ_{∞} , ρ_2 , ρ_3 , ρ_5 , ..., ρ_p , ... are the same non-zero rational numbers, but treated by different norms – ordinary absolute value for subscript ∞ and *p*-adic norms denoted by subscripts *p* = 2, 3, 5, 7, 11, In our case, ρ is a natural number. One can show that the following formula

$$|\rho|_{\infty} |\rho|_2 |\rho|_3 |\rho|_5 \cdots |\rho|_p \cdots = 1 \tag{13}$$

is valid, where product runs over ordinary absolute value and all *p*-adic norms of the same natural (or non-zero rational) number. According to formula (13), all (real and *p*-adic) properties of any non-zero rational number (object) are mutually connected.

In our example with 64 objects, we can consider them as 64 principal ideles. Ordinary absolute value and related usual distance can describe spatial properties of these objects. *p*-Adic norm and the corresponding distance can describe arbitrary many various properties. In particular, by *p*-adic distance one can investigate similarity (and dissimilarity) of some properties between different objects.

The most general mathematical notion for unification of real and *p*-adic numbers is adele, which takes into account field of rational numbers Q with all its non-trivial valuations and all corresponding completions (see, e.g. [10]). Artificial intelligence based on adeles, should be called *adelic intelligence*.

4. Concluding Remarks

In the previous sections we presented some basics of an ultrametric (*p*-adic) approach to language which could be used in design of an artificial intelligence. This *p*-adic approach to an artificial language gives possibility to quantify similarity between words.

As an illustration of ultrametric language, we considered 64 objects presented by 3-digit numbers in base 5 and endowed by *p*-adic properties. Number of these properties can be arbitrary and they are related to *p*-adic norm. Similarity of properties between distinct objects can be expressed by *p*-adic distance. Our example with 64 objects can be generalized to any larger number of words and longer sequences of letters. It is worth noting that these ultrametric spaces of sequences can be also considered as ultrametric networks, where nodes are sequences and links are *p*-adic distances between them.

To compare similarities quantitatively, it is useful to introduce *relative similarity* δ within an ultrametric space U in the following way:

$$\delta(x, y) = (D - d(x, y))/D, \qquad (14)$$

where *D* is maximal distance on space *U*. One can easily see that values of δ are: $0 \le \delta \le 1$. When x = y then $\delta = 1$, and $\delta = 0$ when d(x, y) = D. Thus, relative similarity increases from 0 to 1. This relative similarity can be particularly useful comparing similarities in an ultrametric space U_1 with similarities in another space U_2 . One can also introduce relative dissimilarity $\overline{\delta}$ by formula $\delta + \overline{\delta} = 1$.

For the future investigation of the subject proposed in this paper, we plan to construct a set of objects with appropriate set of natural numbers, and connect *p*-adic properties with some concrete characteristics of given objects.

References

- B. Dragovich and A. Dragovich, A p-adic model of DNA sequence and genetic code, p-Adic Numbers Ultrametric Anal. Appl. 1 (1), 34–41 (2009), [arXiv:q-bio.GN/0607018v1].
- [2] B. Dragovich and A. Dragovich, p-Adic modelling of the genome and the genetic code, Computer J. 53 (4), 432–442 (2010), [arXiv:0707.3043v1 [q-bio.OT]].
- [3] B. Dragovich, p-Adic structure of the genetic code, (2012), [arXiv:1202.2353 [q-bio.OT]].
- [4] B. Dragovich, Genetic code and number theory, Facta Universitatis: Phys. Chem. Techn. 14 (3), 225–241 (2016), [arXiv:0911.4014 [q-bio.OT]].
- [5] B. Dragovich, A. Yu. Khrennokov and N. Ž. Mišić, Ultrametrics in the genetic code and the genome, Appl. Math. Comput. 309, 350–358 (2017), [arXiv:1704.04194 [q-bio.OT]].
- [6] B. Dragovich and N. Ž. Mišić, p-Adic side of the genetic code and the genome, Trends in Bioinformatics: Modeling, Optimization and Computational Problems (Selected works from the BIOMAT Consortium Lectures, Moscow 2017), pp. 75–89 (Springer, 2018).
- [7] A. Khrennikov and S. Kozyrev, *Genetic code on a diadic plane*, Physica A: Stat. Mech. Appl. 381, 265–272 (2007), [arXiv:q-bio/0701007].
 [8] B. Dragovich, A. Yu. Khrennikov, S. V. Kozyrev, I. V. Volovich and E. I. Zelenov, *p-Adic mathematical physics: the first 30 years*, *p*-Adic Numbers Ultrametric Anal. Appl. 9 (2), 87–121 (2017), [arXiv:1705.04758[math-ph]].
- [9] W. H. Schikhof, Ultrametric Calculus: An Introduction to p-Adic Analysis (Cambridge University Press, 1984).
- [10] I. M. Gel'fand, M. I. Graev and I. I. Pyatetski-Shapiro, Representation Theory and Automorphic Functions (Saunders, Philadelphia, 1969).