Applied Mathematics and Computation 309 (2017) 350–358

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

# Ultrametrics in the genetic code and the genome

Branko Dragovich[a,b,*], Andrei Yu. Khrennikov[c], Nataša Ž. Mišić[d]

[a] Institute of Physics, University of Belgrade, Belgrade, Serbia
[b] Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
[c] International Center for Mathematical Modeling in Physics, Engineering, Economics, and Cognitive Science, Linnaeus University, S-35195, Växjö, Sweden
[d] Lola Institute, Kneza Višeslava 70a, Belgrade, Serbia

ARTICLE INFO

*Keywords:*
Ultrametrics
Bioinformation
Genetic code
Ultrametric tree
Ultrametric network
*p*-adic numbers

ABSTRACT

Ultrametric approach to the genetic code and the genome is considered and developed. *p*-Adic degeneracy of the genetic code is pointed out. Ultrametric tree of the codon space is presented. It is shown that codons and amino acids can be treated as *p*-adic ultrametric networks. Ultrametric modification of the Hamming distance is defined and noted how it can be useful. Ultrametric approach with *p*-adic distance is an attractive and promising trend towards investigation of bioinformation.



## 1. Introduction

The choice of mathematical methods in the investigation of physical systems depends on their space and time scale as well as of their complexity. Sometimes standard methods are not sufficient and one has to invent a new advanced method. Biological systems belong to the most complex systems in the nature. In particular, biosystems related to the information processing are very complex and they cannot be completely reduced to the standard physical systems – they are something more than ordinary physical systems and need some new theoretical concepts and mathematical methods to their description and understanding.

It is well known that there is a strong relation between structure and function in living matter. In bioinformation systems we should consider not only physical but also information structure. In the case of physical structure, we use ordinary metrics of Euclidean (or Riemannian) geometry. It is very important to have a metrics which could appropriately describe the structure of a bioinformation as well as similarity (or dissimilarity) between two bioinformation. When we have finite strings (words) of equal length, which are composed of a few different elements (letters), then usually the Hamming distance is used to measure number of positions at which elements (letters) differ. Note that dissimilarity is complementary property to similarity, i.e. less dissimilarity – more similarity, and vice versa. So, one can say that such two strings are more similar as the Hamming distance between them is smaller. However, Hamming distance is not appropriate when informational content of structure elements depends on their place (hierarchy) in the string, e.g. when meaning of elements at the beginning is more important than those at the end. In such case, an ultrametric distance is just an appropriate tool to measure dissimilarity and then bioinformation system can be regarded as an ultrametric space.

Note that an ultrametric space is a metric space in which distance satisfies strong triangle inequality instead of the ordinary one, i.e. $d(x, y) \leq \max\{d(x, z), d(z, y)\}$. As a consequence of this ultrametric inequality, the ultrametric spaces have

* Corresponding author at: Institute of Physics, University of Belgrade, Belgrade, Serbia.
*E-mail address:* dragovich@ipb.ac.rs (B. Dragovich).

http://dx.doi.org/10.1016/j.amc.2017.04.012
0096-3003/

some rather unusual properties, e.g. all triangles are isosceles with one side which cannot be larger than the other two. The Baire metrics between two different words defined to be $2^{-m+1}$, where $m$ is the first position at which the words differ, is an ultrametric distance. Ultrametrics with $p$-adic distances belong to the most elaborated and informative ultrametric spaces. Ultrametrics has natural application in the taxonomy, phylogenesis, genetic code and some complex physical systems [1]. Having many unusual properties, ultrametrics cannot be represented in the Euclidean space, however it can be illustrated in the form of a tree, dendrogram or a fractal.

In this paper we reconsider and further develop $p$-adic approach to the genetic code and the genome introduced in paper [2] and considered in [3–5]. Similar model of the genetic code was considered on diadic plane [6], see also [7]. A dynamical model of the genetic code origin is presented in [8]. In Section 2 some basic properties of ultrametric spaces are presented and illustrated by a few elementary examples with ordinary, the Baire and $p$-adic metrics. Section 3 contains the basic notions of molecular biology including DNA, RNA, codons, amino acids and the genetic code. It also contains the ultrametric trees of codons and amino acids. $p$-Adic structure of the genetic code is described in Section 4, which also contains the ultrametric network aspects of the genetic code. Some $p$-adic ultrametrics of the genome is considered in Section 5. The last section is devoted to conclusion and concluding remarks.

## 2. Ultrametric spaces

The general notion of metric space $(M, d)$ was introduced in 1906 by Fréchet (1878–1973), where $M$ is a set and $d$ is a distance function. Recall that distance $d$ is a real-valued function of any two elements $x, y \in M$ which must satisfy the following properties: $(i)\ d(x,y) \geq 0,\ d(x,y) = 0 \Leftrightarrow x = y,\quad (ii)\ d(x,y) = d(y,x),\quad (iii)\ d(x,y) \leq d(x,z) + d(z,y)$. Property $(iii)$ is called the triangle inequality. An ultrametric space is a metric space where the triangle inequality is replaced by

$$d(x,y) \leq \max\{d(x,z), d(z,y)\}, \tag{1}$$

which is called the strong triangle (also ultrametric or non-Archimedean) inequality. Strong triangle inequality (1) was formulated in 1934 by Hausdorff (1868–1942) and ultrametric space was introduced by Krasner (1912–1985) in 1944.

As a consequence of the ultrametric inequality (1), the ultrametric spaces have many unusual properties. It is worth mentioning some of them.

- *All triangles are isosceles.* This can be easily seen, because any three points $x$, $y$, $z$ can be arranged so that inequality (1) can be rewritten as $d(x,y) \leq d(x,z) = d(z,y)$.
- *There is no partial intersection of the balls. Any point of a ball can be its center. Each ball is both open and closed – clopen ball.* For a proof of these properties of balls, see e.g. [10].

### 2.1. Simple examples of finite ultrametric spaces

Without loss of generality, we are going to present some examples constructed by an alphabet with fixed length $n$ of words endowed with an ultrametric distance. Let $m$ $(m = 1, 2, ..., n)$ be the first position in a pair of words at which letters differ counting from their beginning. Thus $m-1$ is the longest common prefix. Then ultrametrics tell us: the longer common prefix, the closer (more similar) a pair of two words. As illustrative examples, we will take an alphabet of four letters $\mathcal{A} = \{a, b, c, d\}$ and words of length: $n = 1, 2, 3$. Let $W_{k,\,n}(N)$ be a set of words of an alphabet, where $k$ is the number of letters, $n$ is the number of letters in words (length of words) and $N$ is the number of words. Then we have three sets of words: $(i)\, W_{4,\,1}(4)$; $(ii)\, W_{4,\,2}(16)$; $(iii)\, W_{4,\,3}(64)$ (see Table 1). Note that $N = k^n$. In the following we will present ultrametrics of these three different sets with three different distances.

**Ordinary ultrametric distance.** Let us define ordinary ultrametric distance between any two different words $x$ and $y$ as $d(x,y) = n - (m-1)$. It takes $n$ values, i.e. $d(x,y) = 1, 2, ..., n$. Note that one can redefine this distance by scaling it as $d_s(x,y) = \frac{n-m+1}{n}$ and then the scaled distances are between 1 and $\frac{1}{n}$.

- $(i)$ *Case* $W_{4,\,1}(4)$. In this case letters $a$, $b$, $c$, $d$ are words as well. The distance between any two words (letters) is 1, because $n = 1$ and $m = 1$.
- $(ii)$ *Case* $W_{4,\,2}(16)$. Here we have two-letter words (see Table 1). The distance between any two different words $x$ and $y$ is $d(x,y) = 2$ when letters differ at the first position and $d(x,y) = 1$ if letters at the first position are the same $(m = 2)$. Scaling distance is

$$d_s(x,y) = \frac{2-m+1}{2} = \begin{cases} 1, & m = 1 \\ \frac{1}{2}, & m = 2. \end{cases} \tag{2}$$

- $(iii)$ *Case* $W_{4,\,3}(64)$. Now we have three-letter words (see Table 1). Possible values of distance $d(x,y)$ are 1, 2, 3. the corresponding scaling distance is

$$d_s(x,y) = \frac{3-m+1}{3} = \begin{cases} 1, & m = 1 \\ \frac{2}{3}, & m = 2 \\ \frac{1}{3}, & m = 3. \end{cases} \tag{3}$$