p-ADIC SIDE OF THE GENETIC CODE AND THE GENOME

Branko Dragovich
http://www.phy.bg.ac.yu/˜dragovich
dragovich@ipb.ac.rs
Institute of Physics, University of Belgrade
Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia

17th International Symposium on
Mathematical and Computational Biology

30.10 - 3.11. 2017, INM RAS, Moscow, Russia

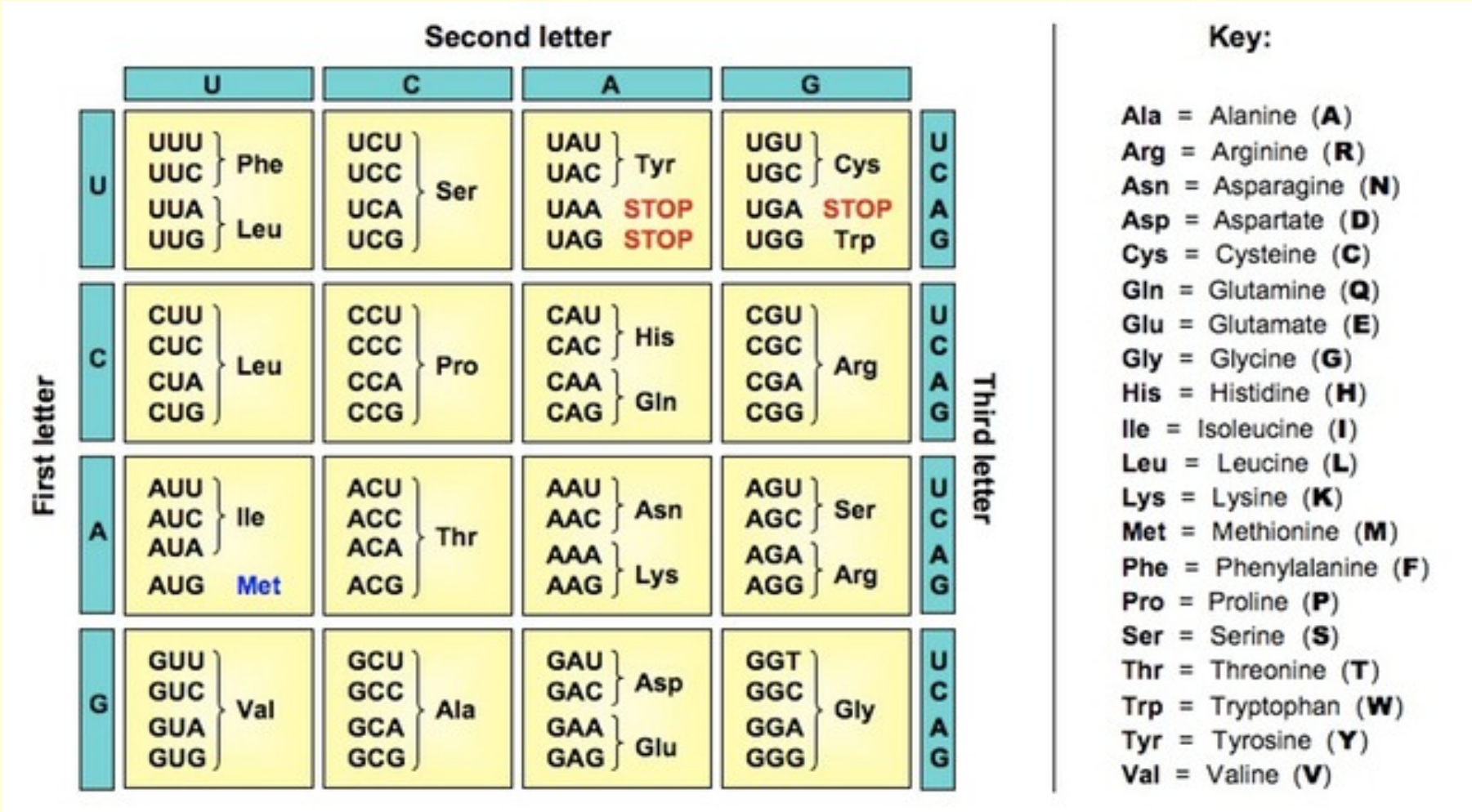
Contents

1. Introduction

- The Genetic Code (GC) is a map from 64 codons onto 20 amino acids + 1 stop signal.
- There are 1.5×10^{84} mapping possibilities.
- Only 31 genetic code in living organisms.
- Codons are building blocks of genes. They are ordered triples of 4 nucleotides (bases): C, A, T(U),G.
- Amino acids (aa) are building blocks of proteins.
- Stop signal (Ter) terminates synthesis of protein.
- In human cells there are two codes: standard and vertebrate mitochondrial code (VMC).
- VMC is simpler than standard code. All codes can be regarded as slight modifications of VMC.

1. Introduction

Table of the standard genetic code



1. Introduction

On modeling of the genetic code

- Why to model genetic code (GC)?
- What are problems in modeling GC?
- Many approaches to model GC:
  - Gamow (1954), Crick (1957), Rumer (1966), ...
  - Hornos and Hornos (1993); Frappat, Sciarrino and Sorba (1998); Forger and Sachse (2000); ...
- *p*-Adic (ultrametric) modeling:
  - Dragovich and Dragovich (2006)
  - Khrennikov and Kozyrev (2007)
  - Bradley (2007)
  - BD ...
  - BD, Khrennikov and Misic (2017)

2. p-Adic distance

- Distance

$$
\begin{aligned}
&(1) \quad d(x,y) \geq 0, \quad d(x,y) = 0 \leftrightarrow x = y \\
&(2) \quad d(x,y) = d(y,x) \\
&(3) \quad d(x,y) \leq d(x,z) + d(z,y)
\end{aligned}
$$

- Ultrametric (non-Archimedean) distance

$$(3.a) \quad d(x,y) \leq \max\{d(x,z), d(z,y)\}$$

- p-adic distance is an example of the ultrametric distance

2. p-Adic distance

- We consider only p-adic distance between (positive) integers
- Let $m, n \in \mathbb{Z}$ and p is a prime number. Then p-adic distance is

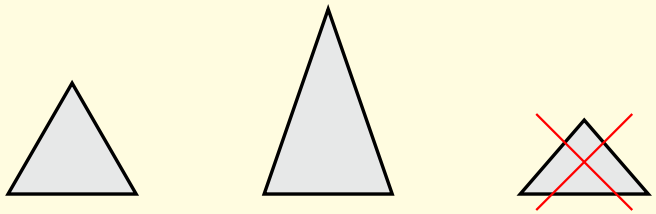
$$d_p(m,n) = |m-n|_p = |p^k q|_p = p^{-k}, \quad k = 0,1,2,3...$$

- $d_p(m,n) \leq 1$ for any $m, n \in \mathbb{Z}$ and any prime number p.

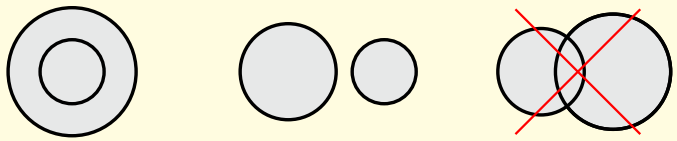
2. p-Adic distance

- Ultrametric distance (F. Hausdorff, 1934) and space (M. Krasner, 1944):

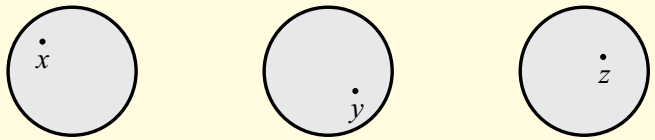
$$(a) \quad d(x,y) \leq \max\{d(x,z), d(z,y)\}$$
$$(b) \quad d(x,y) \leq d(x,z) = d(z,y)$$



All triangles are isosceles.



There is no partial intersection of balls.



Any point of a ball is its center.

3. p-Adic genetic code

- Construction of 64 5-adic numbers

$$C[64] = \{n_0 + n_1 5 + n_2 5^2 \ : \ n_i = 1, 2, 3, 4\}$$
$$n_0 + n_1 5 + n_2 5^2 \equiv n_0 n_1 n_2$$

- Connection of 5-adic $C[64]$ set with set of 64 codons

$$C(Cytosine) = 1, \ A(Adenine) = 2,$$
$$T(Thymine) = U(Uracil) = 3, \ G(Guanine) = 4$$
$$0 = \ absence\, of\, nucleotide$$

3. p-Adic genetic code

Vertebrate Mitochondrial Code

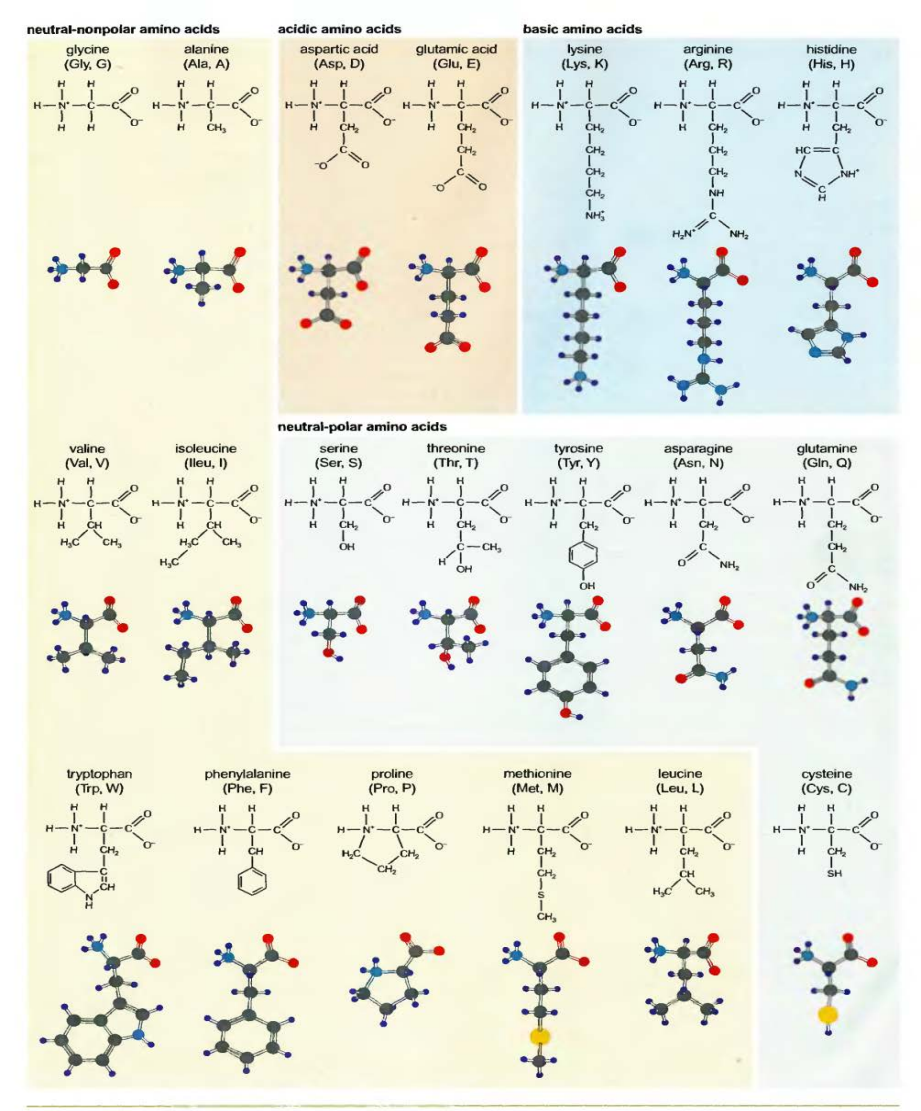
- 64 codons as 32 doublets
- 12 aa coded by single doublets; 6 aa coded by two doublets; 2 aa coded by three doublets; stop signal coded by two doublets.

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

Table : The vertebrate mitochondrial code with p-adic structure.

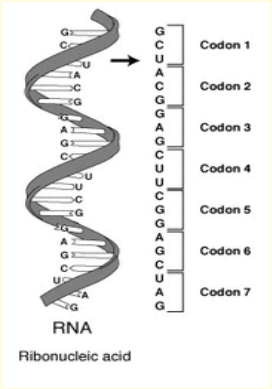
3. *p*-Adic genetic code

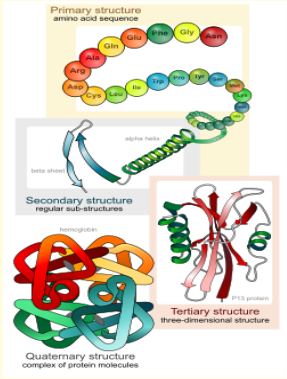
20 standard (canonical) amino acids

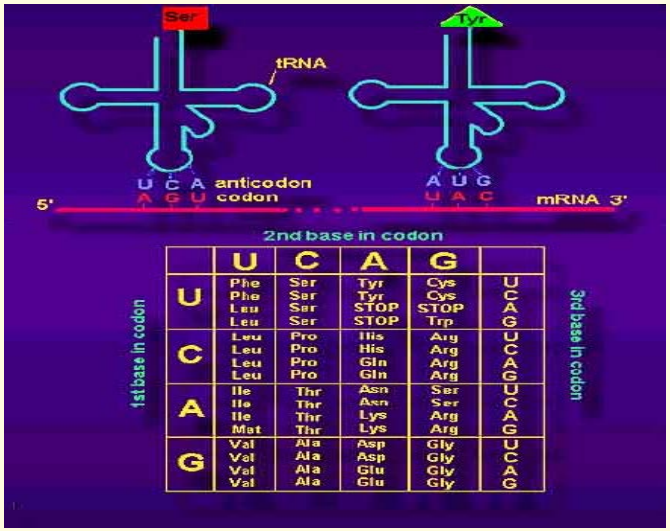


3. *p*-Adic genetic code

codons, amino acids, genetic code







3. p-Adic genetic code

- 5-adic distance between two different codons a and b

$$d_5(a,b) = |a_0 + a_1 5 + a_2 5^2 - (b_0 + b_1 5 + b_2 5^2)|_5$$

- three possibilities:

$$a_0 \neq b_0 \Rightarrow d_5(a,b) = 1$$

$$a_0 = b_0,\ a_1 \neq b_1 \Rightarrow d_5(a,b) = \frac{1}{5}$$

$$a_0 = b_0,\ a_1 = b_1,\ a_2 \neq b_2 \Rightarrow d_5(a,b) = \frac{1}{25}$$

- With respect to the smallest (1/25) 5-adic distance, 64 codons clasterize into 16 quadruplets.

3. p-Adic genetic code

- 2-adic distance between 5-adic quadruplet codons

$$d_5(a,b) = |a_0 + a_1 5 + a_2 5^2 - (b_0 + b_1 5 + b_2 5^2)|_5$$

- Denote codons inside 5-adic quadruplets by a, b, c, d . Then 2-adic distance is:

$$d_2(a,c) = |(3-1)5^2|_2 = \frac{1}{2}$$

$$d_2(b,d) = |(4-2)5^2|_2 = \frac{1}{2}$$

Every quadruplet decays into two 2-adic doublets.

- Now 32 doublets make p-adic basic structure of codon space of 64 elements.

3. *p*-Adic genetic code

- 5-adic distance combined with 2-adic distance correctly describes structure (degeneration) of the set of codons in the vertebrate mitochondrial code (VMC).
- *p*-Adic distance between codons can be in a similar way used for amino acids.

11 CC Pro	21 AC Thr	31 UC Ser	41 GC Ala
12 CA His	22 AA Asn	32 UA Tyr	42 GA Asp
13 CU Leu	23 AU Ile	33 UU Phe	43 GU Val
14 CG Arg	24 AG Ser	34 UG Cys	44 GG Gly

Table : Table of amino acids coded by codons which have pyrimidine at the third position.

3. p-Adic genetic code

11 Pro	12 Thr	13 Ser	14 Ala			
21 His	22 Asn	23 Tyr	24 Asp	212 Gln	222 Lys	242 Glu
31 Leu	32 Ile	33 Phe	34 Val	322 Met		
41 Arg		43 Cys	44 Gly	432 Trp		

Table : 5-Adic distance between amino acids in rows is either $\frac{1}{5}$ or $\frac{1}{25}$ and corresponds to their physicochemical properties.

First row: small size and moderate in hydropathy.
Second row: average size and hydrophilic.
Third row: average size and hydrophobic.
Fourth row: special case of diversity.

3. p-Adic genetic code

Genetic code is an ultrametric network

- Nodes are codons and amino acids. Links are related to 5-adic distances (1/25, 1/5 and 1).
- There are small (d =1/25), intermediate (d=1/5) and large communities (d=1) of codons.
- Genetic code connects two sub networks – network of codons and network of amino acids.

3. p-Adic genetic code

p-Adic evolution of the genetic code

- Primitive genetic code: 4 nucleotides (C, A, U, G) and 4 amino acids (Alanine, Aspartate, Valine, Glycine).
- Dinucleotide genetic code: 16 codons and 16 and 16 amino acids.
- Trinucleotide genetic code: 64 codons and 20 amino acids + stop signal.
- Standard genetic code and other 29 codes can be regarded as slight variations of the vertebrate mitochondrial code.

4. p-Adic similarity between sequences

- 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.
- Hamming distance between these two sequences is $d_H(a,b) = \sum_{i=1}^{n} d(a_i, b_i)$, where $d(a_i, b_i) = 0$ if $a_i = b_i$, and $d(a_i, b_i) = 1$ if $a_i \neq b_i$.
- We introduce p-adically modified Hamming distance: $d_{pH}(a,b) = \sum_{i=1}^{n} d_p(a_i, b_i)$, 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 $d_p(a_i, b_i) < 1$.
- For example, elements a_i and b_i can be nucleotides, codons and amino acids with above assigned natural numbers, and primes $p = 2$ and $p = 5$.

4. p-Adic similarity between sequences

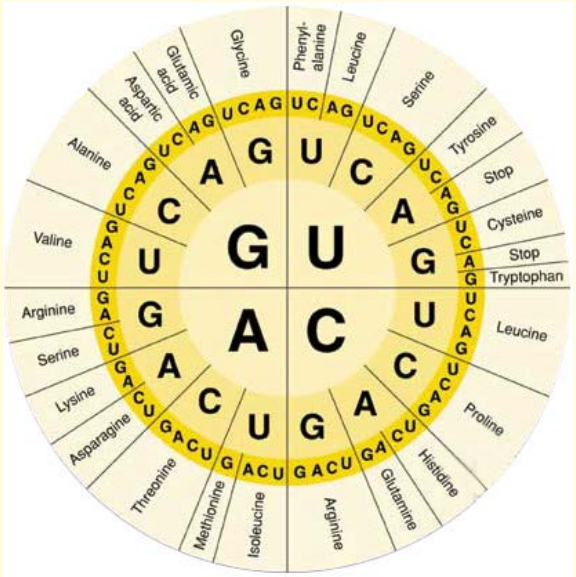
- For sequences as parts of DNA, RNA and proteins, this modified distance (together with Hamming distance) is finer and more informative than Hamming distance itself.
- Example: If $a = a_1\, a_2\, a_3 = (111)(412)(443)$ and $b = b_1\, b_2\, b_3 = (113)(414)(441)$ then the corresponding Hamming distance is $d_H(a, b) = 3$, while p-adic modified ones are $d_{5H}(a, b) = \frac{3}{25}$ and $d_{2H}(a, b) = \frac{3}{2}$. Now suppose that we do not know exactly these two sequences a and b, but we have information on their distances. If we would know only the Hamming distance we could not conclude at which three positions of related codons nucleotides differ. However, taking 5-adic and 2-adic modified Hamming distances together, it follows that codon differences are at the third position of nucleotides and that sequences a and b code the same sequence of amino acids, in fact the sequence *ProAlaGlu*.

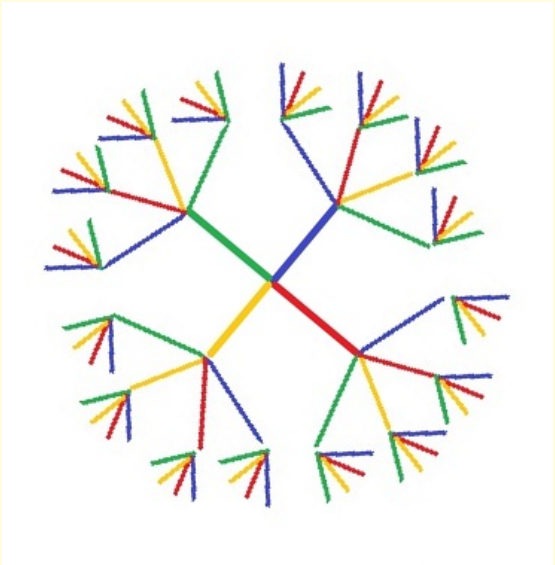
5. Concluding remarks

- p-Adic distance is a simple and adequate tool for description of the genetic code: similarity between codons, similarity between amino acids, and connection between codons and amino acids.
- p-Adic distance is also useful for investigation of bioinformation (sequence between nucleotides or between amino acids) similarity.
- Similar in structure – similar in function!

5. Concluding remarks

- Ultrametric tree of codons





5. Concluding remarks

- Genetic code can be considered as dictionary between two biomolecular languages – 1) 4 letters (nucleotides) and 2) 20 letters (amino acids) languages.
- Codons – three-letter words.
- Proteins – multi-letter words

5. Concluding remarks

Applications of *p*-adic (ultrametric) analysis.

1. In some very short-distance systems
   - *p*-adic strings
   - space-time geometry at the Planck scale
   - quantum systems
2. In some very complex systems
   - spin glasses
   - conformational dynamics of proteins
3. In some information systems
   - genetic code
   - bioinformation
   - taxonomy
   - phylogenetics
   - linguistics
   - sequences of symbols

Some references

- B. Dragovich and A. Dragovich, “A *p*-adic model of DNA sequence and genetic code,” *p*-Adic Numbers Ultrametric Anal. Appl. **1** (1) (2009) 34–41. arXiv:q-bio.GN/0607018v1.
- B. Dragovich and A. Dragovich, “*p*-Adic modelling of the genome and the genetic code,” Computer Journal **53** (4) (2010) 432–442. arXiv:0707.3043v1 [q-bio.OT].
- B. Dragovich, “Genetic code and number theory,” (2009). arXiv:0911.4014 [q-bio.OT]
- B. Dragovich, “*p*-Adic structure of the genetic code,” (2012). arXiv:1202.2353 [q-bio.OT].
- A. Khrennikov and S. Kozyrev, ”Genetic code on a diadic plane,” Physica A: Stat. Mech. Appl. **381** (2007) 265–272. arXiv:q-bio/0701007.
- B. Dragovich, A. Khrennikov and N.Z. Misic, “Ultrametrics in the genetic code and the genome,” Applied Mathematics and Computation **309** (2017) 350–358.

THANK YOU FOR YOUR ATTENTION!