p-ADIC GENETIC CODE AND ULTRAMETRIC BIOINFORMATION

Branko Dragovich
http://www.phy.bg.ac.yu/˜dragovich
dragovich@ipb.ac.rs
Institute of Physics, Mathematical Institute SASA, Belgrade

6th International Conference
p-Adic Mathematical Physics and its Applications

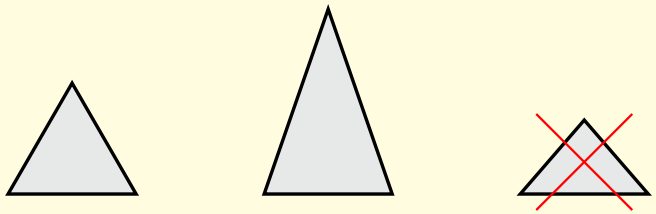
23.10 - 27.10. 2017, CINVESTAV, Mexico City, Mexico

Contents

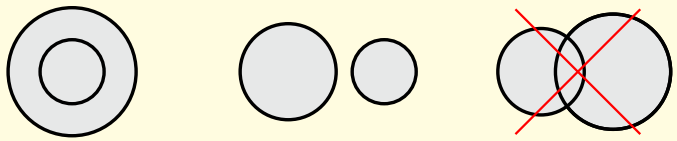
1. INTRODUCTION: ultrametric space

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

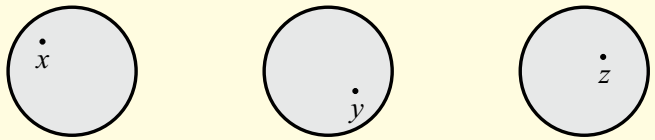
$$(a) \quad d(x,y) \leq \max\{d(x,z), d(z,y)\}$$
$$(b) \quad d(x,y) \leq d(x,z) = d(z,y)$$



All triangles are isosceles.



There is no partial intersection of balls.



Any point of a ball is its center.

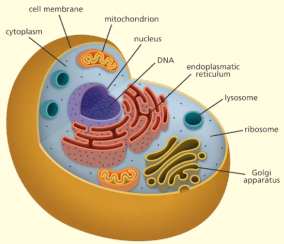
1. INTRODUCTION: ultrametric space

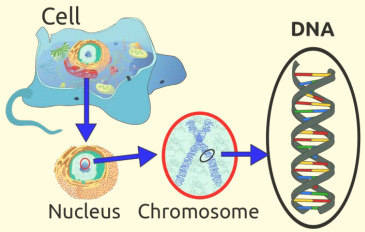
– Applications?

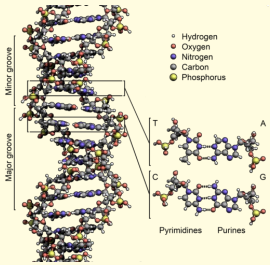
1. In some very short-distance systems
   - p-adic strings
   - space-time geometry at the Planck scale
   - quantum systems
2. In some very complex systems
   - spin glasses
   - protein dynamics
3. In some information systems
   - genetic code
   - bioinformation
   - taxonomy
   - phylogenetics
 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

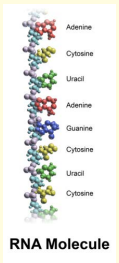
DNA, RNA, nucleotides, codons, ...

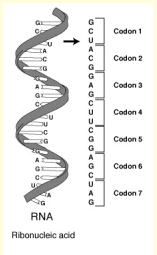
- Codons are ordered triples of four nucleotides (bases, letters): C = Cytosine, A = Adenine, T= Thymine (U = Uracil) and G = Guanine. $4 \times 4 \times 4 = 64$ codons

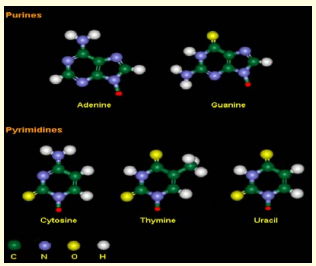




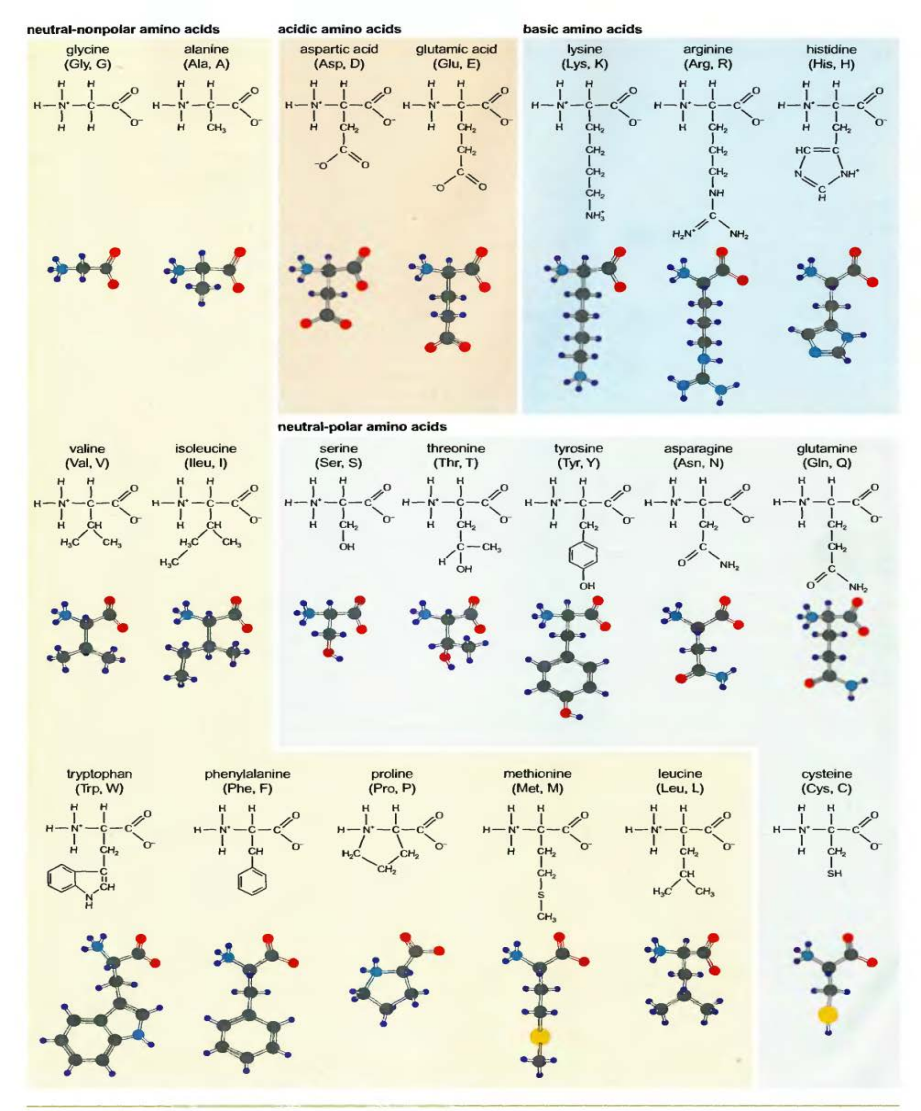




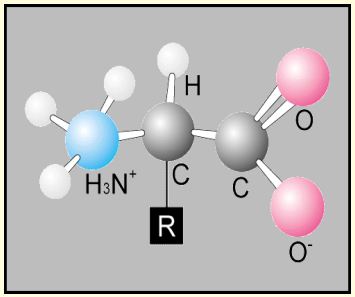


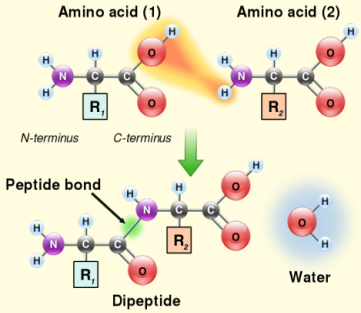


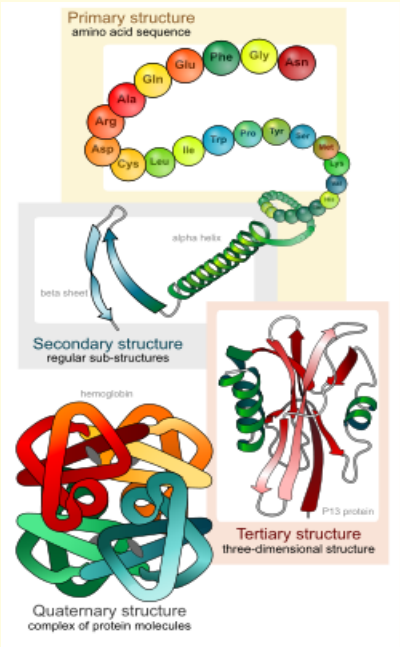
20 standard (canonical) amino acids

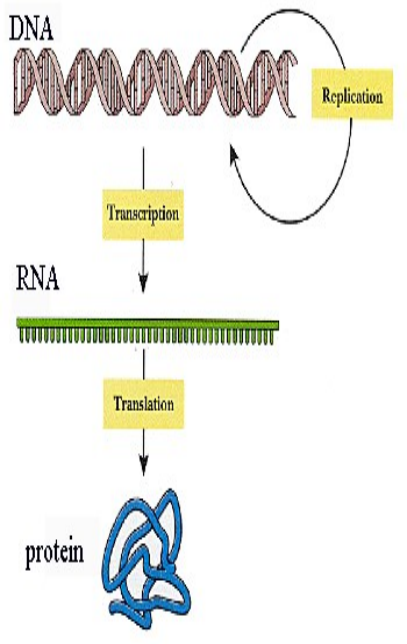


peptides, proteins, from DNA to proteins



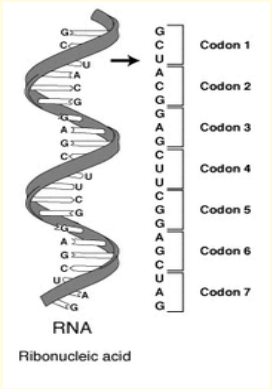


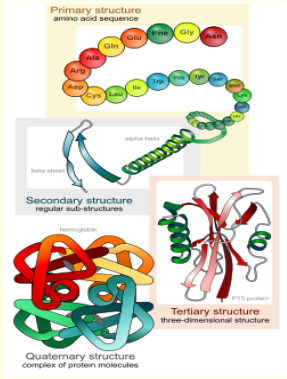


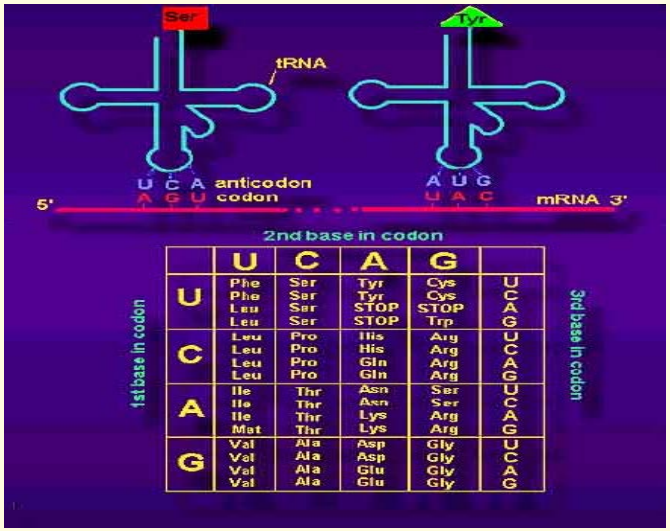


3. ON GENETIC CODE

ribosomes, transport RNA, standard genetic code

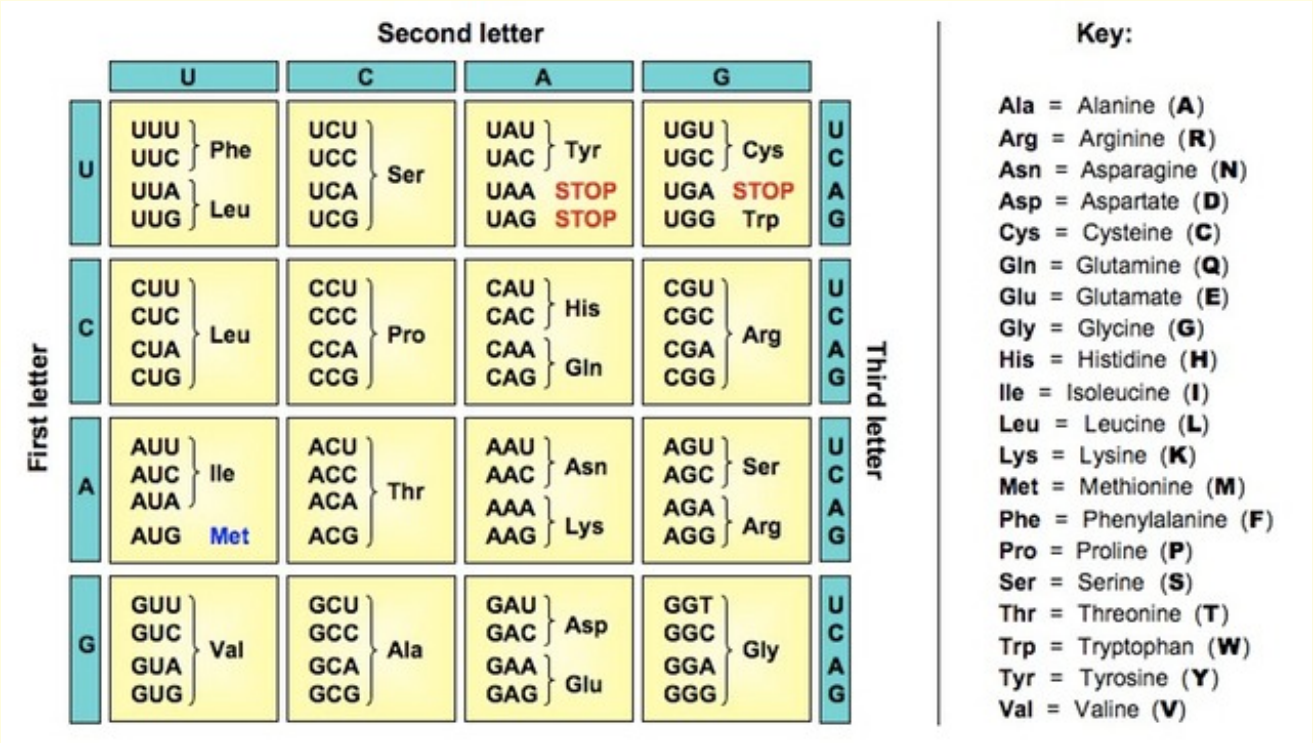






3. ON GENETIC CODE

- The genetic code is a map from 64 codons onto 20 amino acids + 1 stop signal.
- There are 1.5×10^{84} mappings.
- Only 31 genetic code in living organisms.
- In human cells 3×10^{9} base pairs in DNA. Only 1.5% of DNA codes proteins.
- 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.



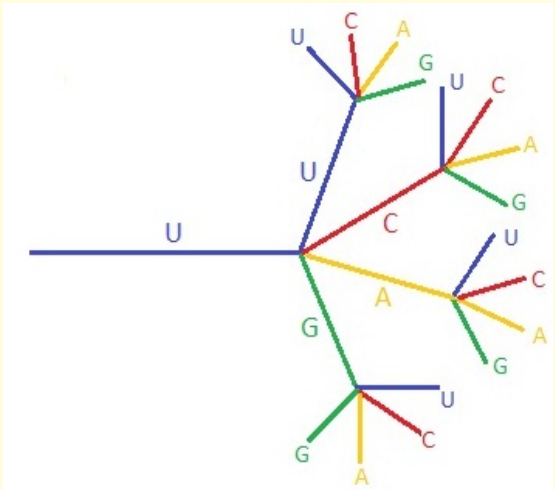
3. ON GENETIC CODE: modeling of genetic code

- 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)

4. ON p-ADIC GENETIC CODE

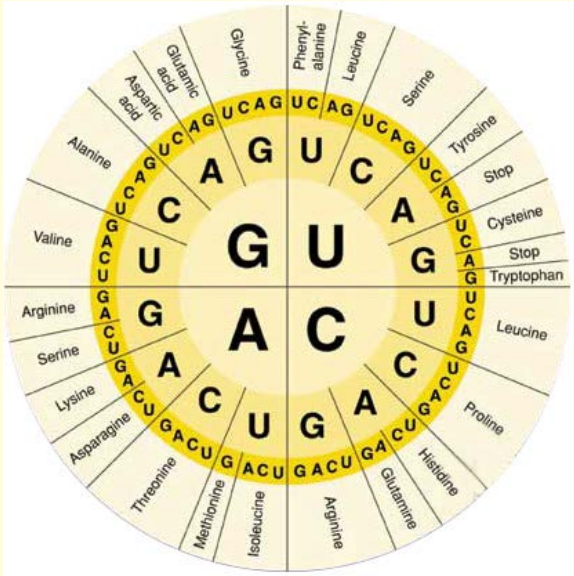
codons as a tree

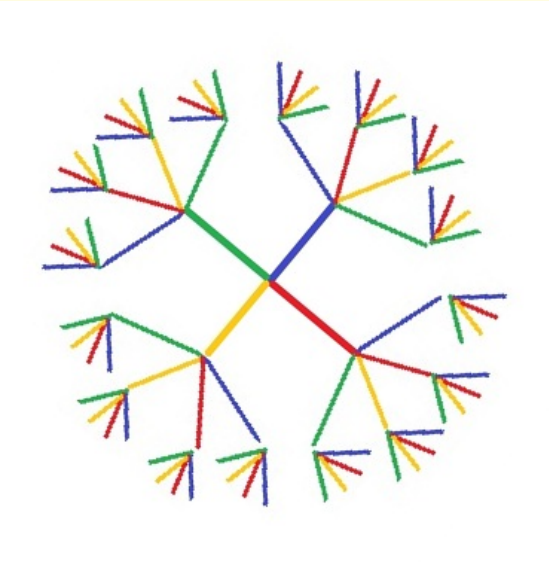
First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
A	AUU Ile, AUC Ile, AUA Ile, AUG Met	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn, AAA Lys, AAG Lys	AGU Ser, AGC Ser, AGA Arg, AGG Arg	U C A G
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp, GAA Glu, GAG Glu	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G



4. ON p-ADIC GENETIC CODE

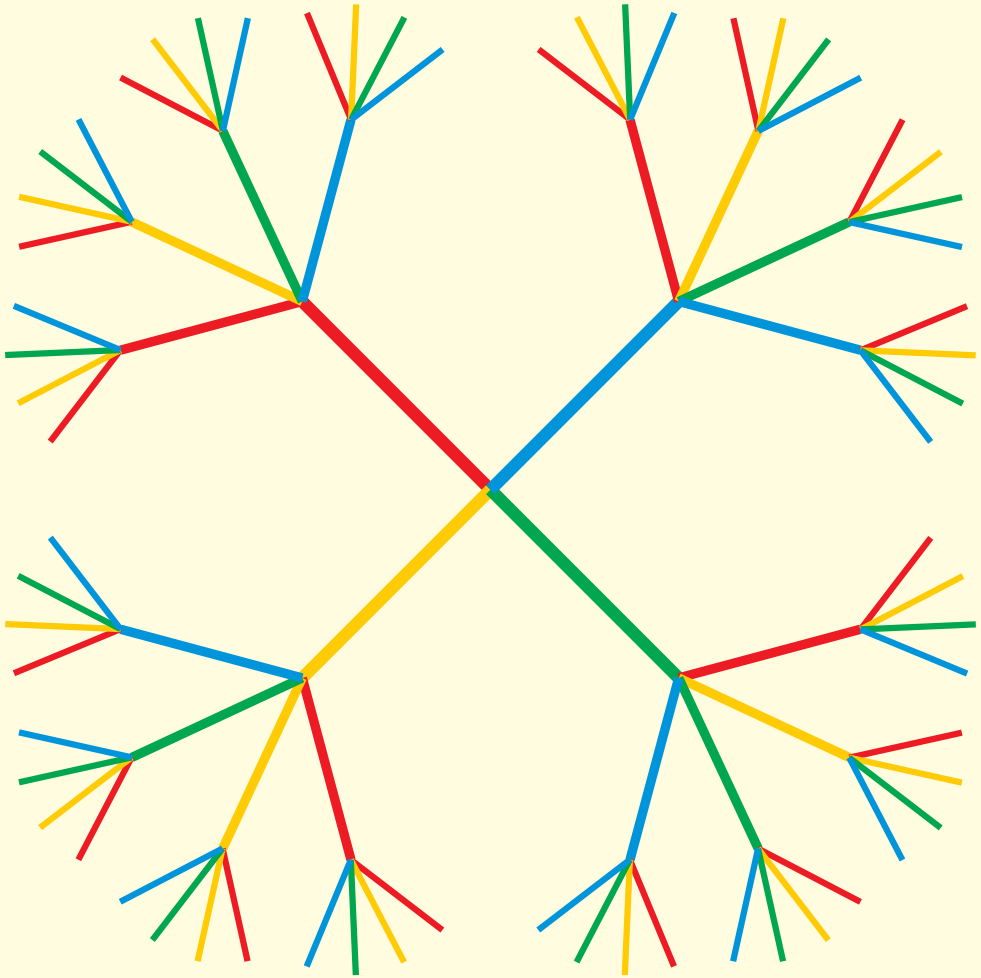
ultrametric codons tree





4. ON p-ADIC GENETIC CODE

- Ultrametricity of codon space



Ultrametric Tree of the Genetic Code

4. ON p-ADIC GENETIC CODE

- p-Adic space of codons

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

- C (Cytosine) = 1, A (Adenine) =2, TThymine = U (Uracil) = 3, G (Guanine) = 4

  0 = absence of nucleotide

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

Vertebral Mitochondrial Code

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	33 3 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

4. ON 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.

4. ON 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 to two 2-adic doublets.

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

5. ON BIOINFORMATION AND SIMILARITY

- Bioinformation? Any sequence of nucleotides or amino acids
- Similarity? Similarity between two sequences.
- Why? Similar in structure – similar in function!

5. ON BIOINFORMATION AND SIMILARITY

- Let $a = a_1\, a_2\, \cdots a_n$ and $b = b_1\, b_2\, \cdots b_n$ be two strings of equal length.
- Hamming distance between these two strings 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 in the following way: $d_{pH}(a,b) = \sum_{i=1}^{n} d_p(a_i, b_i)$, 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 $d_p(a_i, b_i) < 1$.
- In the case of strings as parts of DNA, RNA and proteins, this modified distance is finer and should be more appropriate than Hamming distance itself. 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 $p = 5$.

6. CONCLUDING REMARKS

- The Genetic Code has p-adic ultrametric structure and is very simple and evident application of the p-adic distance.
- It describes degeneracy of the codon space.
- One can also introduce p-adic ultrametric structure of 20 amino acids.
- Genetic code is an ultrametric network. Network of codons – small, intermediate and large community.
- Ultrametric (p-adic) approach to evolution of the genetic code is also considered.
- Ultrametric (p-adic) similarity is important for study bioinformation.

7. Main references

- B. Dragovich, A. Dragovich, “A p-adic model of DNA sequence and genetic code,” p-Adic Numbers Ultrametric Anal. Appl. 1 (1) (209) 34–41. arXiv:q-bio.GN/0607018v1.
- B. Dragovich, A. Dragovich, “p-Adic modelling of the genome and the genetic code,” Computer J. 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, 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,” Appl. Math. Comp. 309 (2017) 350–358.