ON *p*-ADIC GENETIC CODE AND BIO-INFORMATION

Branko Dragovich
dragovich@ipb.ac.rs

Institute of Physics, University of Belgrade
Belgrade, Serbia

International Workshop on
p-Adic Methods for Modeling of Complex Systems
ZiF, Bielefeld University
15.04 - 19. 04. 2013
Bielefeld – Germany

Contents

1. Introduction
2. DNA, RNA, Proteins, and Genetic Code
3. 5-Adic space of 64 elements
4. *p*-Adic structure of the genetic code
5. *p*-Adically modified the Hamming distance
6. Conclusions

1. Introduction

- Francis Crick (1916–2004), who together with James Watson discovered double helicoidal structure of DNA, in 1953 announced:

> We have discovered the secret of life.

- However, the life has still many secrets.
- The genetic code is also a secret of life in the following sense:

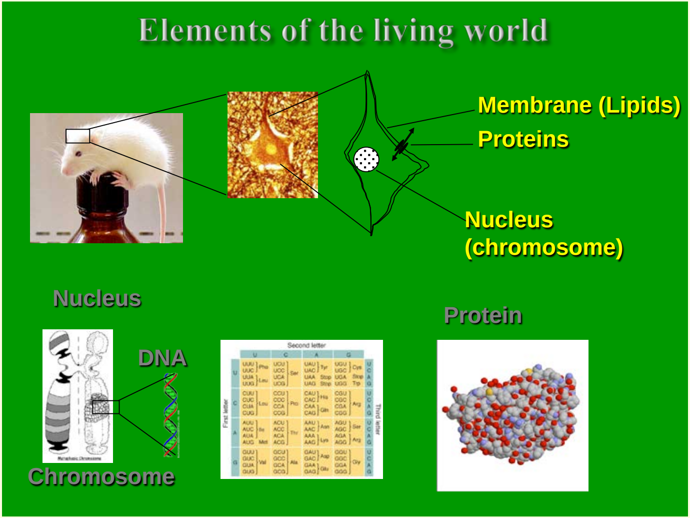
> We mainly know how to describe structure of the genetic code, but we don't know origin and evolution of its structure.

- I shall present here p-adic description of the genetic code structure.

1. Introduction

- Dragovich, B. and Dragovich, A.: A *p*-Adic Model of DNA Sequence and Genetic Code. *p*-Adic Numbers, Ultrametric Analysis and Applications, Vol. 1, No. 1, 34–41. (2009). [arXiv:q-bio.GN/0607018v1]
- Dragovich, B. and Dragovich, A.: *p*-Adic Modelling of the Genome and the Genetic Code. The Computer Journal, Vol. 53, No. 4, 432–442. (2010). [arXiv:0707.3043v1 [q-bio.OT]]
- Dragovich, B.: *p*-Adic Structure of the Genetic Code. NeuroQuantology, Vol. 9, No. 4, 716–727. (2011). [arXiv:1202.2353v1 [q-bio.OT]]
- Khrennikov A. and Kozyrev, S.: Genetic Code on a Diadic Plane. Physica A: Stat. Mech. Appl., Vol. 381, 265–272. (2007). [arXiv:q-bio/0701007]

1. DNA, RNA, Proteins, and Genetic Code

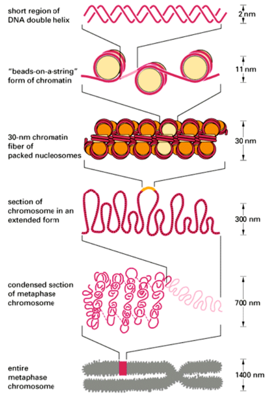


Elements of the living world

Second letter

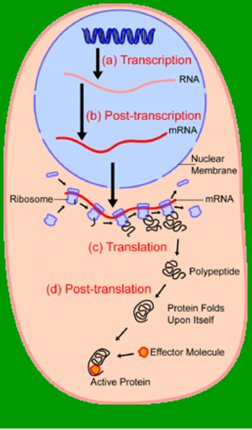
First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA Stop	UGA Stop	A
	UUG	UCG	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C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

Complex Structure of DNA

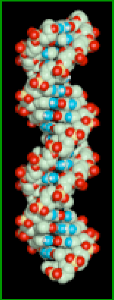

short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in an extended form
300 nm
condensed section of metaphase chromosome
700 nm
entire metaphase chromosome
1400 nm

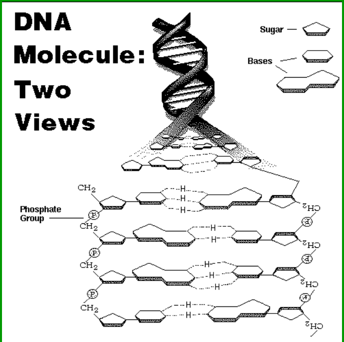
Excellent try, my dear Watson, but . . .
could you bring me a mirror, old boy?
Simon '98

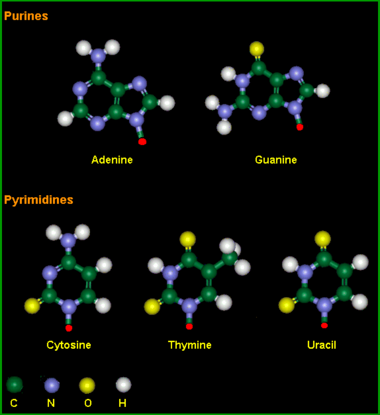
From DNA to proteins

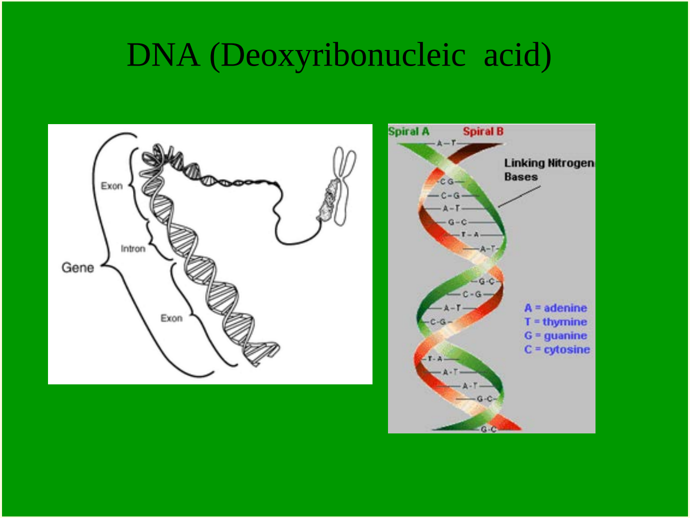


DNA (1953)



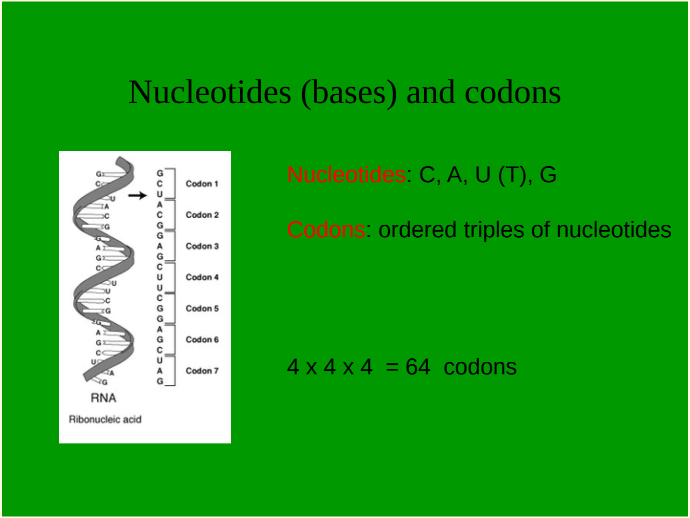



Purines
Adenine
Guanine
Pyrimidines
Cytosine
Thymine
Uracil
C
N
O
H



DNA (Deoxyribonucleic acid)

Ribonucleic acid


G
C
U
A
C
G
G
A
G
C
U
U
C
G
G
A
G
C
U
A
G
Codon 1
Codon 2
Codon 3
Codon 4
Codon 5
Codon 6
Codon 7
RNA
Ribonucleic acid

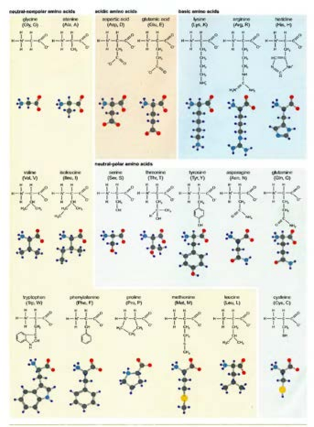
Nucleotides (bases) and codons

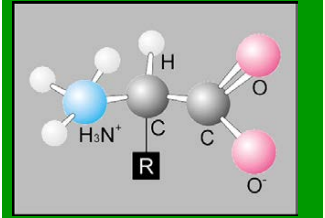
Nucleotides: C, A, U (T), G

Codons: ordered triples of nucleotides

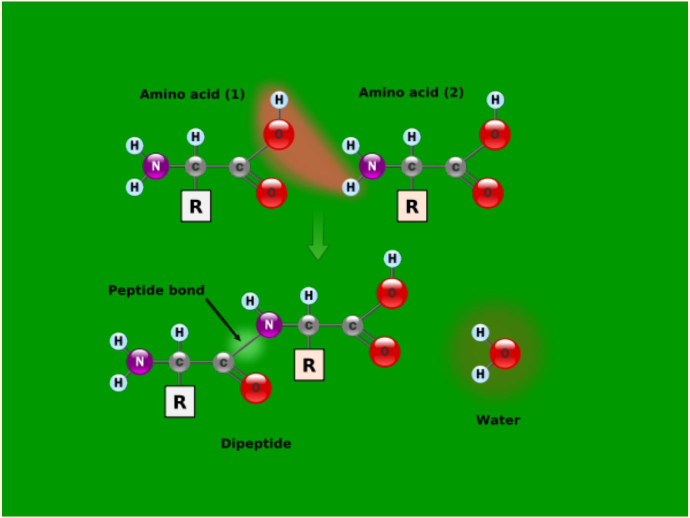
4 x 4 x 4 = 64 codons

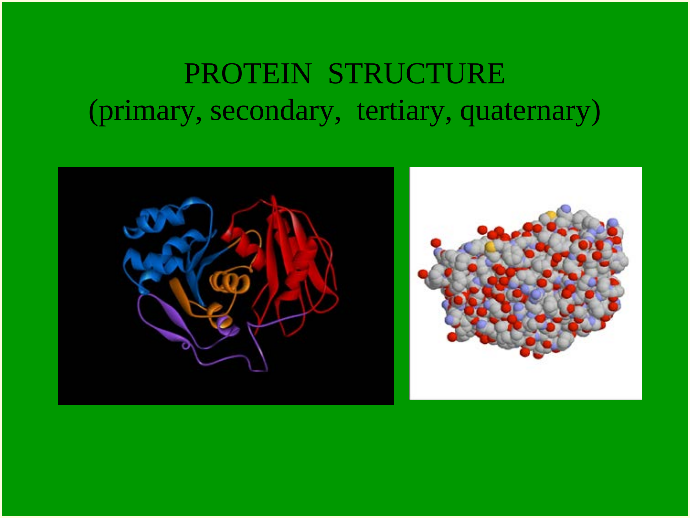
Amino acids





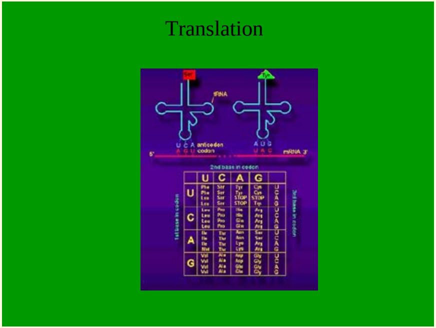
20 standard a.a. +
Selenocysteine
Pyrrolysine


Amino acid (1)
Amino acid (2)
H
O
N
C
R
Peptide bond
Dipeptide
Water



PROTEIN STRUCTURE
(primary, secondary, tertiary, quaternary)

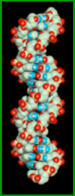
Protein syntheses in the ribosomes

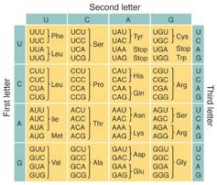


Translation

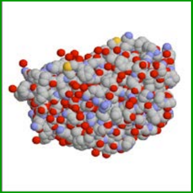
	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

From DNA to proteins
using the GENETIC CODE





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

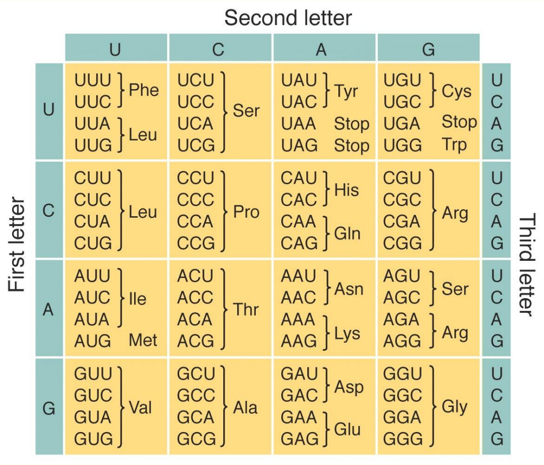


2. Genetic Code

- The standard genetic code was deciphered about 1965.
- The genetic code (GC) is connection between 64 codons, which are building blocks of the genes, and 20 amino acids, which are building blocks of the proteins.
- Mathematically: The genetic code is a map of 64 elements onto 21 element.
- In addition to coding amino acids, a few codons code stop signal, which is at the end of genes and terminates process of protein synthesis.
- Codons are ordered triples composed of C, A, U (T) and G nucleotides.
- Each codon presents an information which controls use of one of the 20 standard amino acids or stop signal in synthesis of proteins.

2. Genetic Code

- The GC is usually presented by a table.

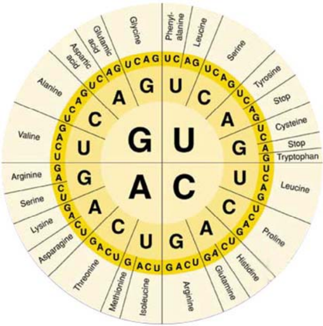


	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile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

2. Genetic Code

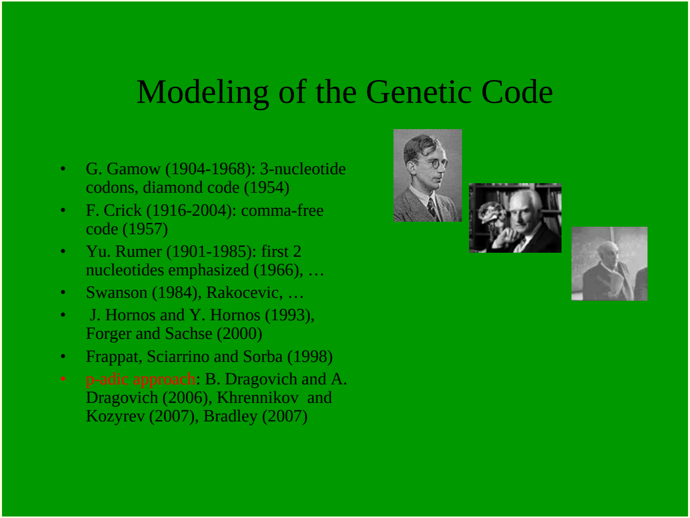
- Another table for the GC.

Another representation of the GENETIC CODE



2. Genetic Code

- Why to investigate modelling of the genetic code?
- From mathematical point of view, the GC is a mapping of a set of 64 elements onto a set of 21 elements.
- There is in principle a huge number (more than 10^{80}) of possible mappings, but the genetic code is one definite mapping with a few (about 20) slight modifications.
- For modelling of the GC, the main problem is to find the corresponding structure of the space of 64 and 20 (or 21) elements.
- It will be demonstrated here that the set of 64 codons, and 20 amino acids, has p-adic structure, where $p = 5$ and 2.



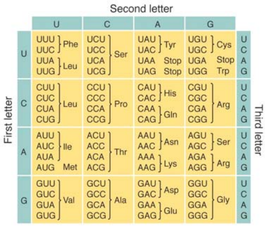
Modeling of the Genetic Code

- G. Gamow (1904-1968): 3-nucleotide codons, diamond code (1954)
- F. Crick (1916-2004): comma-free code (1957)
- Yu. Rumer (1901-1985): first 2 nucleotides emphasized (1966), …
- Swanson (1984), Rakocevic, …
- J. Hornos and Y. Hornos (1993), Forger and Sachse (2000)
- Frappat, Sciarrino and Sorba (1998)
- p-adic approach: B. Dragovich and A. Dragovich (2006), Khrennikov and Kozyrev (2007), Bradley (2007)

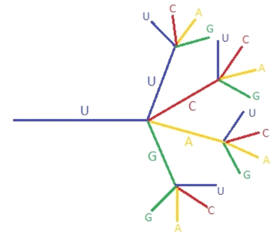
2. Genetic Code

- The GC in the form of an ultrametric tree.

p-Adic approach:Ultrametric (Tree) Structure of Codons

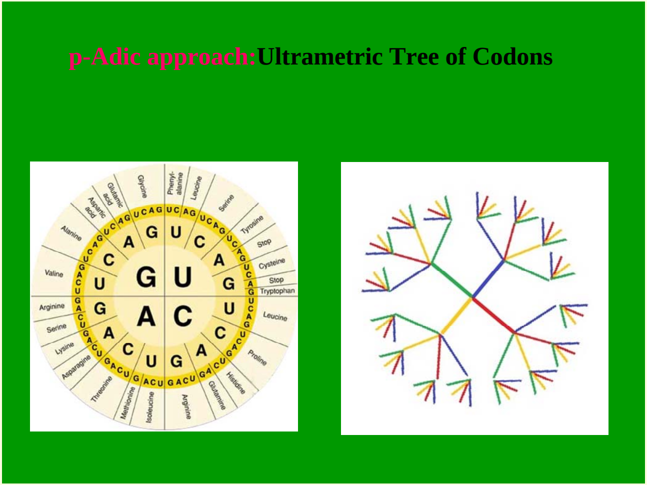


	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile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



2. Genetic Code

- The GC in the form of an ultrametric tree.



p-Adic approach:Ultrametric Tree of Codons

3. 5-Adic Space of 64 elements

- We introduce 64 natural numbers using 5-adic expansion
- From *p*-adic mathematics we use only *p*-adic distance
- We use operation which is change of digits, but without summation and multiplication
- We identify 5-adic space of 64 elements with 64 codons, and call it *p*-adic codon space and consider it as a simple illustration of bio-information system with *p*-adic structure.

3. 5-Adic Space of 64 elements

- We introduce the following set of 64 natural numbers:

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

- 5-adic distance

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

  can be:

  1. $d_5(a, b) = 1$ if $a_0 \neq b_0$
  2. $d_5(a, b) = \frac{1}{5}$ if $a_0 = b_0$ and $a_1 \neq b_1$
  3. $d_5(a, b) = \frac{1}{25}$ if $a_0 = b_0$, $a_1 = b_1$ and $a_3 \neq b_3$

- With respect to the smallest 5-adic distance, there is clustering of $C[64]$ into quadruplets.

111 112 113 114	211 212 213 214	311 312 313 314	411 412 413 414
121 122 123 124	221 222 223 224	321 322 323 324	421 422 423 424
131 132 133 134	231 232 233 234	331 332 333 334	431 432 433 434
141 142 143 144	241 242 243 244	341 342 343 344	441 442 443 444

3. 5-Adic Space of 64 elements

- 2-adic distance between elements within 5-adic quadruplets
- denote elements of quadruplets by a, b, c, d, respectively. Then
  1. $d_2(a,c) = |(3-1)\cdot 5^2|_2 = |2|_2\,|25|_2 = |2|_2 = \frac{1}{2}$
  2. $d_2(b,d) = |(4-2)\cdot 5^2|_2 = |2|_2\,|25|_2 = |2|_2 = \frac{1}{2}$
  3. in other cases $d_2(\cdot,\cdot) = 1$.
- by 5-adic and 2-adic distance we get 32 doublets which can be connected with the genetic code of human (vertebrate) mitochondria by identification

> C (Cytosine) =1, A (Adenine) = 2, U (Uracil) = T (Thymine) = 3, G (Guanine) = 4 .

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

4. *p*-Adic structure of the genetic code

- The genetic code of human (vertebrate) mitochondria can be viewed as the basic one
- Amino acids are coded by one, two or three codon doublets
- Standard genetic code can be obtained from this code by:
    1. 234 AUA: Met → Ile
    2. 242 AGA and 244 AGG : Ter → Arg
    3. 342 UGA : Trp → Ter
- Other (about 20)known versions of the genetic code in some living systems can be also obtained from this one by its slight modification.
- These modifications are like broken symmetry in physics – reality is realization of some broken symmetries.

4. *p*-Adic structure of the genetic code

Table : 20 standard amino acids with assigned 5-adic numbers become close to codon doublets.

11 Proline	21 Threonine	31 Serine	41 Alanine
12 Histidine	22 Asparagine	32 Tyrosine	42 Aspartate
13 Leucine	23 Isoleucine	33 Phenylalanine	43 Valine
14 Arginine	24 Lysine	34 Cysteine	44 Glycine
1 Glutamine	2 Methionine	3 Tryptophan	4 Glutamate

$$x = x_0 + x_1\,5 \equiv x_0\,x_1\,, \qquad x = x_0\,, \quad x_i = 1,2,3,4.$$

4. p-Adically modified the Hamming distance

- 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 $p = 5$.

5. Conclusions

- Codons which are *p*-adically closest (the most similar) code the same amino acid.
- 5-adic and 2-adic distance describe structure of the codon space in the form of 32 doublets.
- To each doublet corresponds one amino acid in the vertebrate mitochondrial code.
- All versions of the genetic code can be viewed as slight modifications of the vertebrate mitochondrial code.
- *p*-Adic Hamming distance is introduced, which should be useful in alignments of strings of codons and strings of amino acids.