ULTRAMETRICITY IN BIOINFORMATION SYSTEMS

Branko Dragovich

Institute of Physics, Belgrade, Serbia

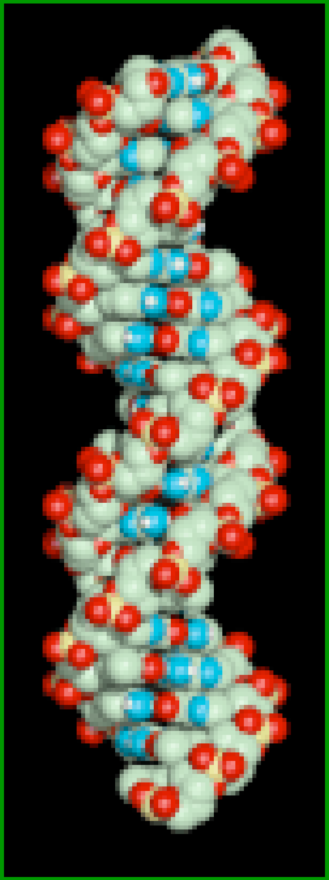
http://ipb.ac.rs/~dragovich

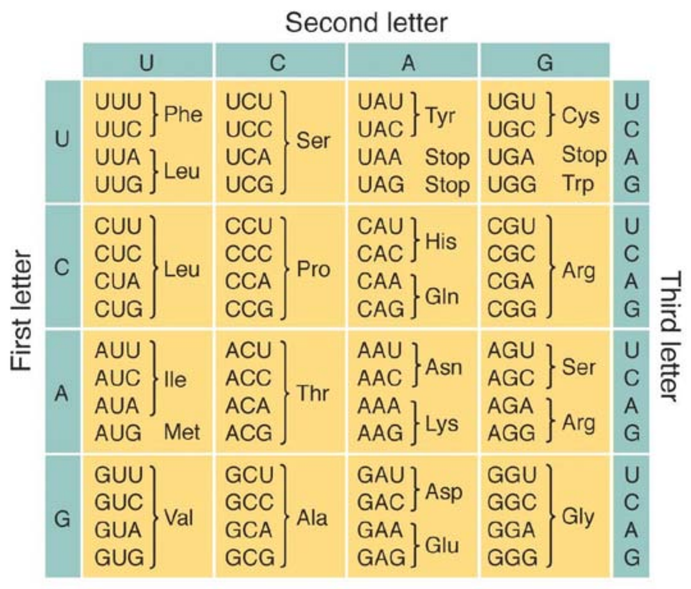
dragovich@ipb.ac.rs

TABIS2013

Belgrade, 21.09.2013

From DNA to proteins
using the GENETIC CODE

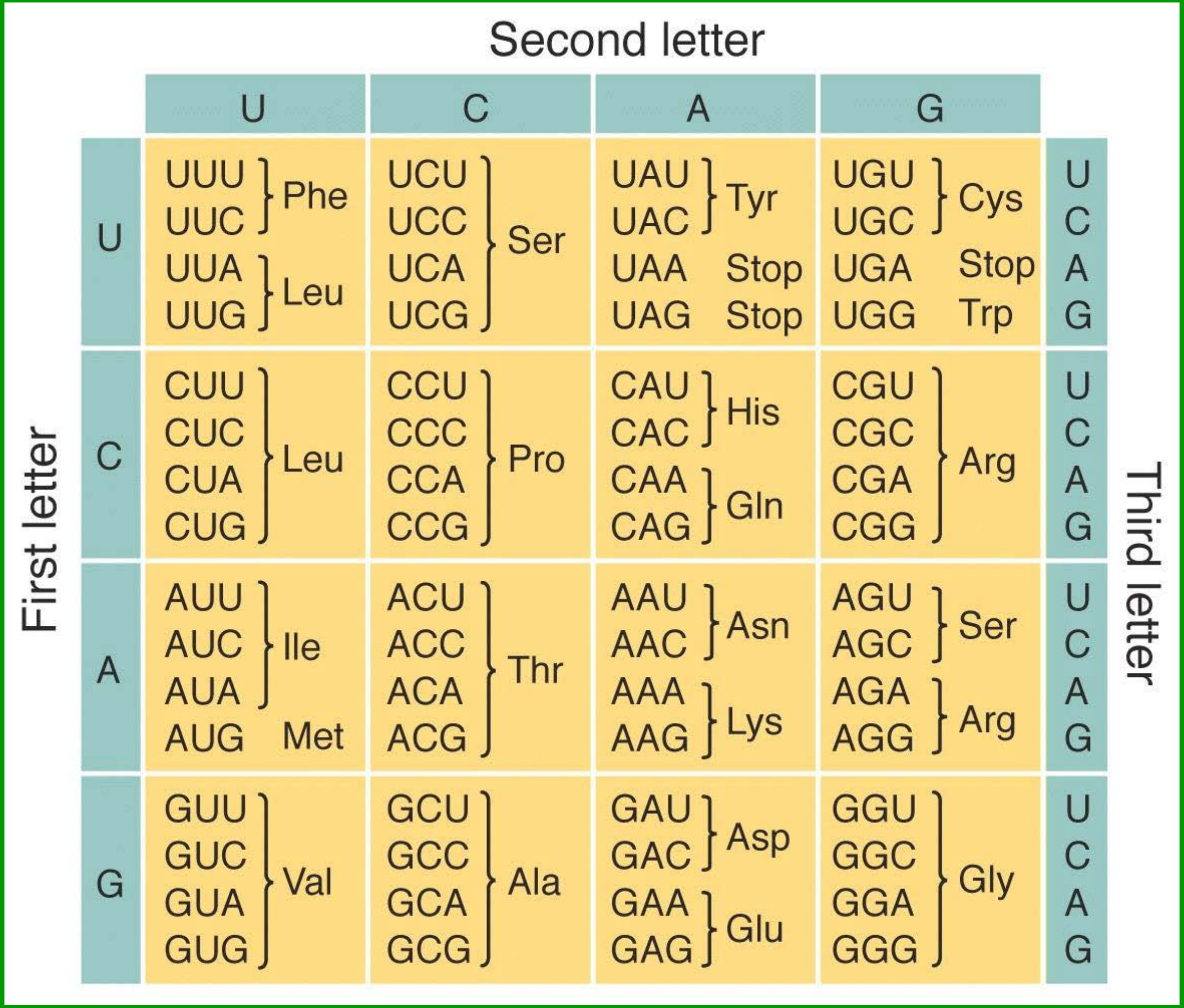




Second letter

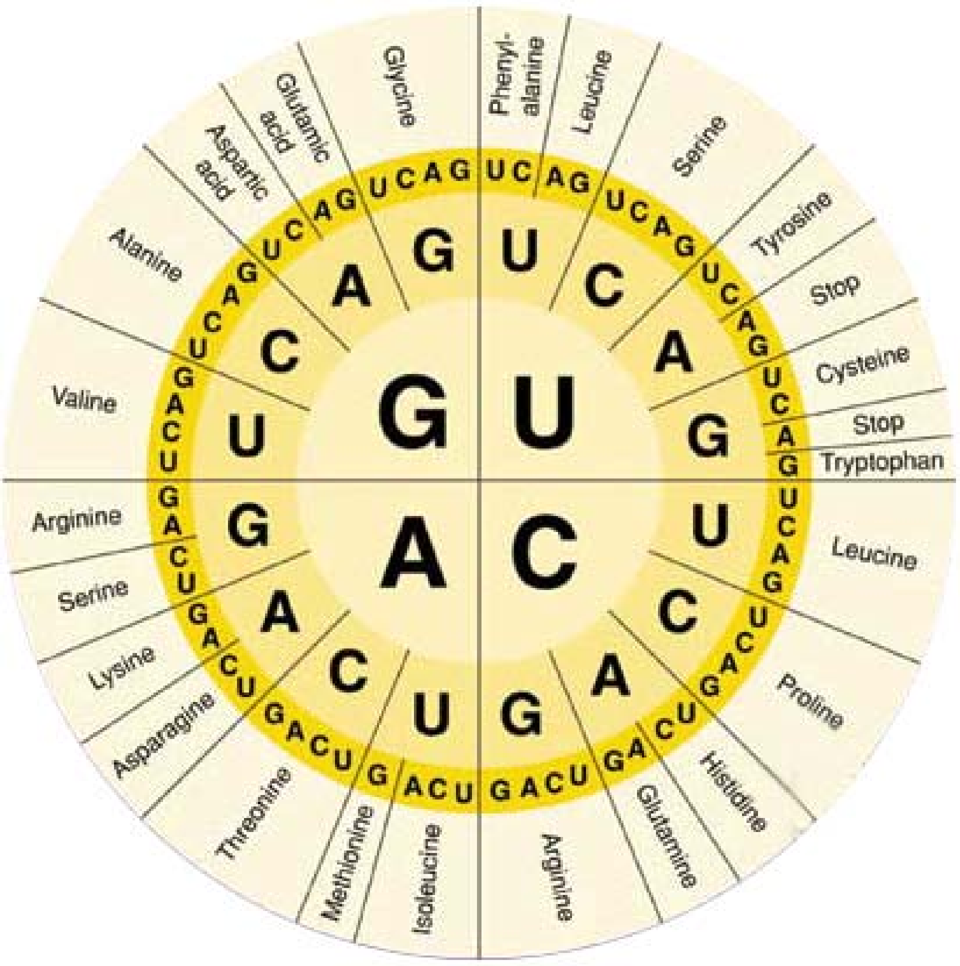
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	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

Table of the GENETIC CODE



	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

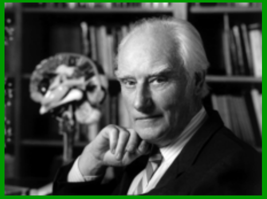
Another representation of the GENETIC CODE



Motivations to Model the Genetic Code

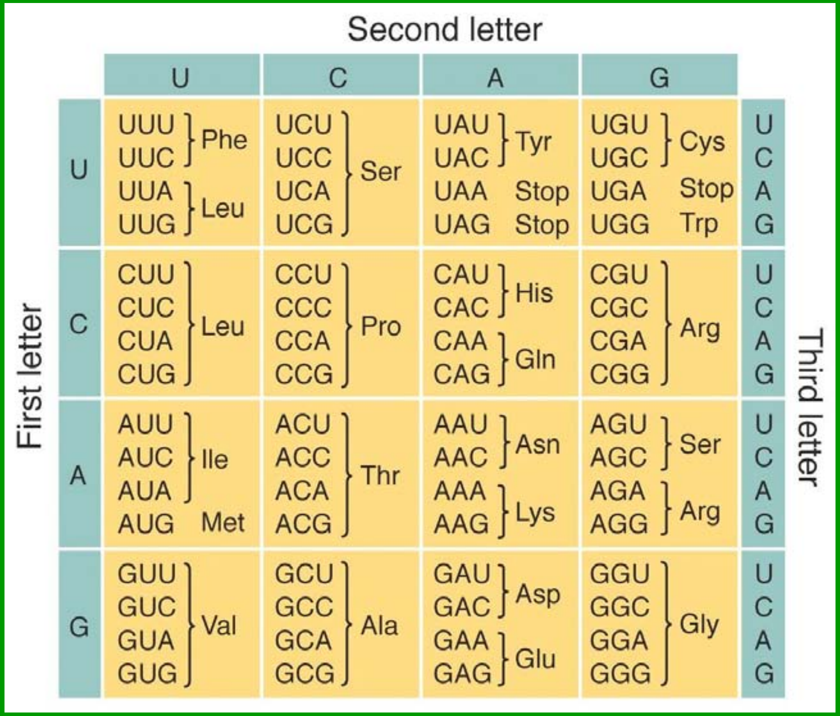
- Genetic code is a map from 64 elements (codons) to the set of 21 elements (20 amino acids and 1 stop signal).
- There is a huge number of possibilities for genetic coding, but exists in living organisms only one genetic code and about 20 its slight modifications.
- What is the rule behind the Genetic Code?
- What is structure of the space of codons?
- What was origin of the genetic code? What was its evolution so far? What is its possible evolution in the future?

Modeling of the Genetic Code

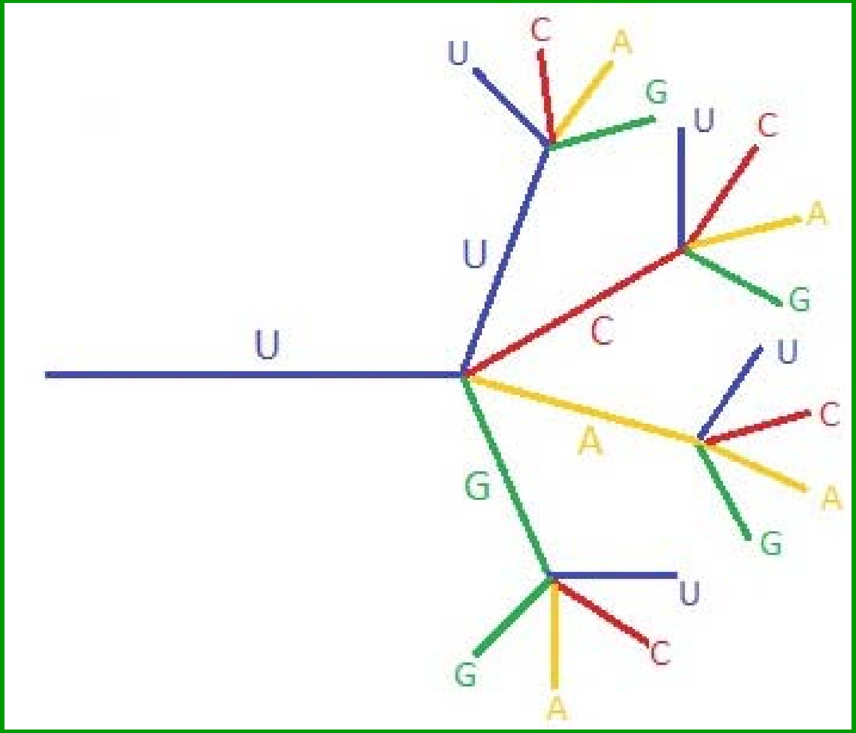


- G. Gamow (1904-1968): 3-nucleotide codons, diamond code (1954)
- F. Crick (1916-2004): comma-free code (1957)
- Yu. Rumer (1901-1985): first 2 nucleotides emphasized (1966), …
- Scherbak, Rakocevic, Misic,…
- J. Hornos and Y. Hornos (1993), Forger and Sachse (2000)
- Frappat, Sciarrino and Sorba (1998)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

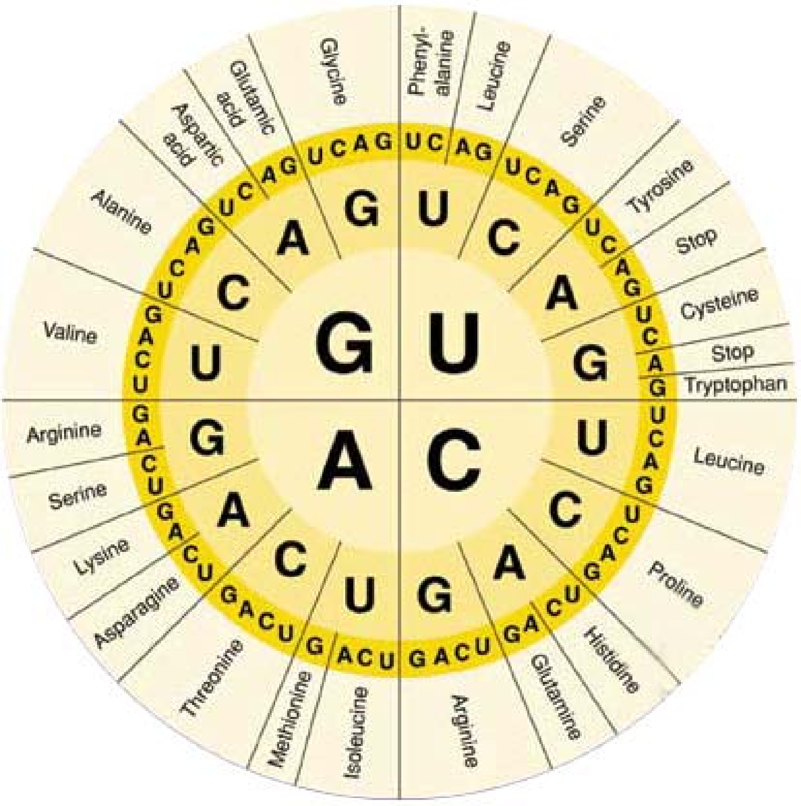
p-Adic approach:Ultrametric (Tree) Structure of Codons

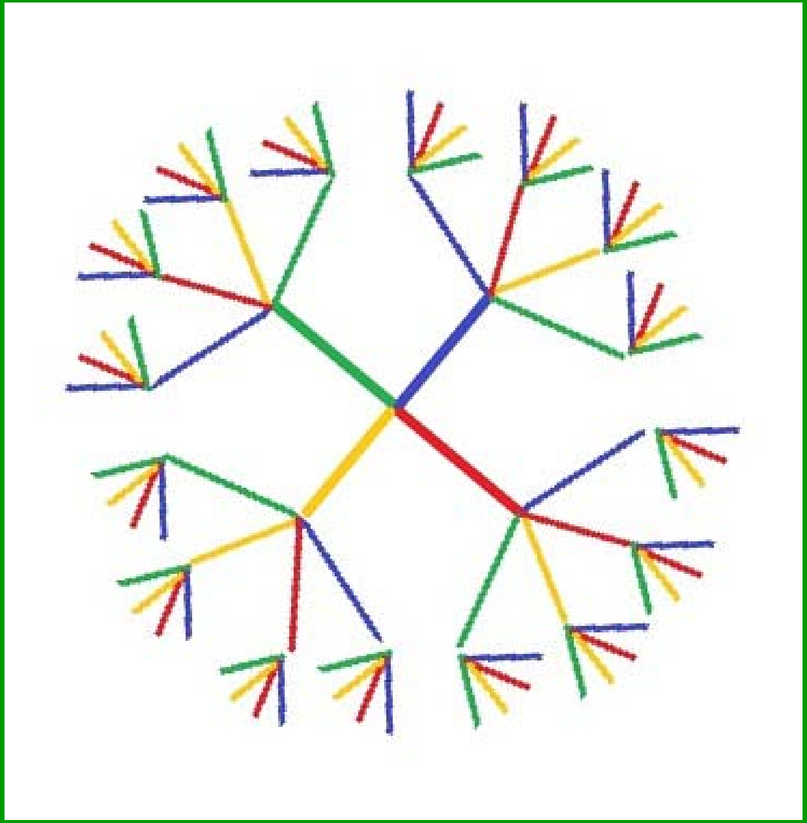


	Second letter				
First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA Stop	UGA Stop	A
	UUG	UCG	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	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



p-Adic approach:Ultrametric Tree of Codons





p-Adic approach: ultrametric spaces

- metric space: Maurice Frechet (1906)
- word “ultrametric” by Marc Krasner (1944)
- p-adic spaces: Kurt Hensel (1897)
- ultrametric distance in Taxonomi: Karl Linne (1735)
- hierachical structures in proteins
- genetic code

Kurt Hensel
(1861 – 1941)

p-Adic Distance Between Integers

- Let a and b be two integers, and let a-b =c =p^k m, where p is a prime number, k is not negative integer, and m is not divisible by p. Then p-adic distance is:

$$d_p(a,b) = | a - b |_p = | c |_p = | p^k m |_p = p^{-k}$$

which is ultrametric, i.e.

$$d_p(x,y) \le \max\left\{d_p(x,z), d_p(z,y)\right\} \le d_p(x,z) + d_p(z,y)$$

p-Adic Space of Codons

5-adic natural numbers with three digits different from zero:

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

$$n_0 + n_1 5 + n_2 5^2 \equiv n_0 n_1 n_2$$

C (cytosine) = 1, A (adenine) = 2,
T (thymine) = U (uracil) = 3, G (guanine) = 4
(0 = absence of nucleotide)

5-adic distance between codons a and b is

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

When codons a and b are different, there are 3 possibilities:

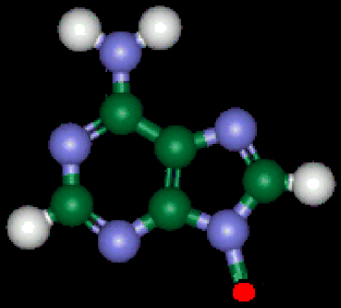
$$d_5(a,b) = 1 \Leftarrow a_0 \neq b_0$$

$$d_5(a,b) = \tfrac{1}{5} \Leftarrow a_0 = b_0, a_1 = b_1$$

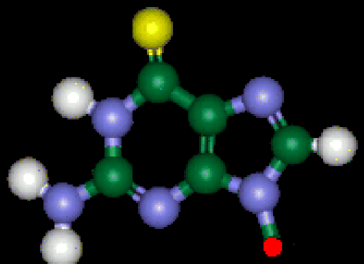
$$d_5(a,b) = \tfrac{1}{25} \Leftarrow a_0 = b_0, a_1 = b_1, a_2 \neq b_2$$

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

Purines

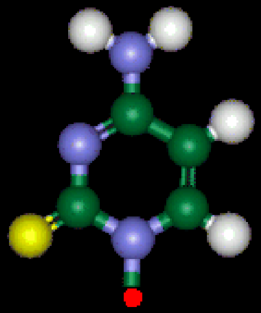


Adenine

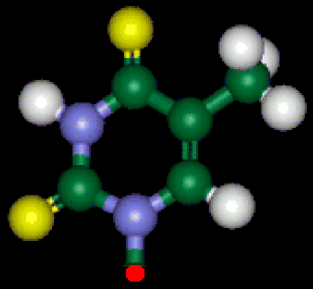


Guanine

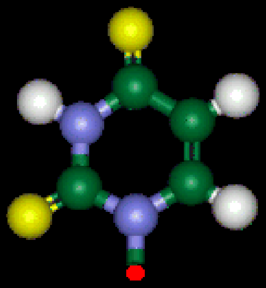
Pyrimidines



Cytosine



Thymine



Uracil

● C ● N ● O ● H

2-adic distance between 5-adic quadruplet codons

- Denote codons inside 5-adic quadruplet by

a, b, c, d . Then 2-adic distance is:

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

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

Now every quadruplet decays into two 2-adic doublets. **These 32 doublets make basic structure of space of 64 codons**.

Code of Vertebral Mitochondria

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

5-adic distances: 1/25, 1/5, 1

2-adic distances: 1/2, 1

p-Adic Properties of the Vertebral Mitochondrial Code

- T-symmetry: doublets-doublets and quadruplets-quadruplets invariance
- 5-Adic distance gives quadruplets
- 2-Adic distance inside quadruplets gives doublets
- Degeneration of the genetic code has p-adic structure
- p-Adic degeneracy principle: *Codons code amino acids and stop signals by doublets which are result of combined 5-adic and 2-adic distances*
- Modern assignment of codon doublets to particular amino acids is a result of coevolution of the genetic code and amino acids: single nucleotide code – 4 amino acids, dinucleotide code – 16 amino acids, trinucleotide code 20 amino acids.
- Other (19) codes may be regarded as slight modifications of the Code of Vertebral Mitochondria

Examples of p-Adic Codon Spaces

$$\Gamma_p[(p-1)^m]$$

- 1-nucleotide codon space: p=5, m =1
- 2-nucleotide codon space: p=5, m =2
- 3-nucleotide codon space: p=5, m =3

Possible evolution of codon space:

$$\Gamma_5(4) \to \Gamma_5(4^2) \to \Gamma_5(4^3) = \Gamma_5(64)$$

CONCLUSIONS

- Codon space has ultrametric structure.
- p-Adic distance plays important role in the genetic code.
- 4 nucleotides are structural elements of 5-adic space of 64 codons.
- Codons clasterize according 5-adic and 2-adic distances.
- Degeneration of the Genetic Code has p-adic structure.
- This p-adic approach can be extended to some other aspects of genomics and proteomics.