Heuristic methods for analyzing some properties of the genetic code

I. TRENCHEV, I. PAVLOVA,
Department of Information Systems and Technologies
University of Library Studies and Information Technologies
c. Sofia, pk 1784, bul. "Tsarigradsko Shosse" № 119
BULGARIA
i.trenchev@unibit.bg

Abstract**.** Many articles have been presented attempting to explain the origin of the genetic code. The codes which have the properties of genetic code but do not collide in nature are called theoretical genic codes. The cardinality of this set is around 1084. In this paper we show a new optimization according to simple mutation. In this pater, we present resistance of the genetic code against simple mutation and this factor is responsible for the evolution of the genetic code.

Key-Words: - Genetic code, optimality, Heuristic methods

1 Introduction

The contemporary genetic code (CGC) is the mapping of 64 three-letter codons to 20 amino-acids and a stop signal. It is to be expected that the contemporary genetic code is structure ensures maximum resistance to mutation effects. Let consider the genetic code as a system, for storage, transmission, execution and regulation of the information encoded in the genes. So it is worthy to analyze the resistance against the effects of mutations, which are the equivalent of noise or errors inherent to all information systems [1-8, 11-15].

All codes which have the properties of the genetic code, but don't collide in nature are called theoretical genetic codes (TGC)[27, 30].

All 61 codons coded for 20 amino acids, there must be degeneracy in code, and some amino acids AA are encoded by several triplets. For example there are two different codons: UUU and UUC, which order for amino acid Phenylalanine. The set of all this codons which code same amino acid is named synonymous set. Any specific codon coded only one amino acid. The genetic code is unambiguous [1-6]. A deciphering of the information which is encoded in the DNA doesn't involve any overlapping. The process begins with a specific codon and during the all reading of the information in DNA there isn't any punctuation between codons. It stops when a nonsense codon is reached. In all live nature the genetic code is a same (universal) [7 -14].

It can be shown that different theoretical codes built from a fixed number of triplets resist to the effects of mutations differently, depending on the relative positions of their codons in the 64 possible divisions.

All previous works analyzed the optimality of groups of triplets translated into the same amino acid. Never has been measured correctly decoded with probability the resistance of the whole genetic code to the creation of non synonym mutants. The present paper deals with the notion of resistance of the whole genetic code to mutation effects. Here we defined a new optimization principle and observed a good correspondence between the contemporary genetic code and the theoretical considerations[14 - 23].

* Corresponding author: trenchev@swu.bg

2 Methods

If we want to understand a real properties of the nature and specters of life we must analyze CGC. In the present work this problem is solved us comparing of the CGC with set of all TGC. If the criteria is chosen a correctly we will characterize CGC. In the previous paper we investigate a CGC. We describe the set of all TGC us a convex polytope. Our polytope description gives a possibility to analyze all TGC. This description will give us ground to characterize more deeply the properties of the CGC [12, 15, 16, 21-23].

Let A***, C, G***, and ***U*** are letters form set ***A ={A, C, G, U}*** and let ***M*** is a set of all three letters words over set A. The cardinality of ***M*** is 64 (4^3). Also we defined a set a ={a_1, a_2, …, a_{23}} because the number of all amino acids are 20 and three stop codons. Let L ={L_1, L_2, …, L_{23}} is a partition of M with following properties[27, 30]:

$$\bigcup L_i = M \quad L_i \neq \varnothing \quad i = 1,...,23 \quad L_i \bigcap L_j = \varnothing; \quad i, j = 1,...,23$$

Our main goal is to solve and analyze the solutions of the following general discrete optimization task:

$$F\ (K(X),\ parameters,\ ) \rightarrow extremum.$$

with conditions $K(X) \in K(L)$,

We are looking for such a TCG that is the minimum or maximum of the criterion of optimality F.
We use the criteria of optimality and obtained this TGC [16]:

$$F(x_{ij}) = \sum_{i=1}^{20}\left(\sum_{m=1}^{64} x_{im}\right)^{p_i}\left(\sum_{j=1}^{64}\sum_{k=1}^{64} P\left(x_{ij}, x'_{ik}\right) M\left(a(x_{ij}), a'(x'_{ik})\right)\right) \rightarrow \max$$

conditions

$$x = \left\{x_{ij}, i=1,..23\, j=1,..64 \sum_{i=1}^{23}\sum_{j=1}^{64} x_{ij} = 64 \sum_{j=1}^{64} x_{ij} \geq 1\ i=1,..20 \sum_{j=1}^{64} x_{ij} = 1, \quad i=21..23\right\}$$

where
p_i are probability of occurrence of amino acids in average protein,

$P\left(x_{ij}, x'_{ik}\right)$ - probability of replacing the codon by CGC with codons TGC

$M\left(a(x_{ij}), a'(x'_{ik})\right)$ - mutation matrix for replacing an amino acid encoded by a contemporary genetic code and one of the theoretical genetic code,

Table 1. Probability of occurrence of amino acids in average protein [16]

Amino acids	**Biological spieces**		
AK	Archea	Bacteria	Eukaryotes
Ala	0.0785	0.0808	0.0648
Arg	0.0592	0.0499	0.0524
Asp	0.0547	0.0506	0.0531
Asn	0.034	0.0463	0.0476
Cys	0.0089	0.01	0.0186
Glu	0.0779	0.0635	0.0664
Gln	0.019	0.0389	0.0428
Gly	0.0749	0.067	0.0588
His	0.017	0.0207	0.0241
Ile	0.0759	0.0705	0.0548
Leu	0.0965	0.1052	0.0935
Lys	0.0604	0.0643	0.063
Met	0.0249	0.0219	0.0233
Phe	0.04	0.0457	0.042
Pro	0.0443	0.0399	0.0515
Ser	0.0593	0.0618	0.085
Thr	0.0477	0.0515	0.0557
Trp	0.0103	0.011	0.0113
Tyr	0.0368	0.0323	0.0303
Val	0.0797	0.0687	0.0609

Table 2. Theoretical genetic code obtained using the probabilities for the appearance of Amino acid in Archea.

First position	Second position				Third position
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stp	Stp	A
	Leu	Ser	Stp	Val	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U

A	Ile	Thr	Asn	Ser	C
	Ile	Trp	Lys	Met	A
	Leu	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Table 3. Theoretical genetic code obtained using the probabilities for the appearance of Amino acid in Bacteria.

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

Table 4. The number of codons of the Contemporary genetic code (CGC) and the various theoretical genetic codes (TGC) derived from Monte Carlo simulations

Amino acids	CGC	TGC		
		Archea	Bacteria	Eukaryotes
Ala	**4**	**4**	**5**	**4**
Arg	**6**	**6**	**6**	**5**
Asp	**2**	**2**	**2**	**2**
Asn	**2**	**2**	**2**	**2**
Cys	**2**	**2**	**2**	**2**
Glu	**2**	**2**	**2**	**2**
Gln	**2**	**2**	**2**	**2**
Gly	**4**	**4**	**4**	**4**
His	**2**	**2**	**2**	**2**
Ile	**3**	**3**	**3**	**3**
Leu	**6**	**7**	**7**	**7**
Lys	**2**	**2**	**2**	**2**
Met	**1**	**1**	**1**	**1**
Phe	**2**	**2**	**2**	**2**
Pro	**4**	**4**	**4**	**4**
Ser	**6**	**7**	**6**	**6**
Thr	**4**	**3**	**3**	**3**
Trp	**1**	**1**	**1**	**1**
Tyr	**2**	**2**	**2**	**2**
Val	**4**	**3**	**3**	**5**
Stp	**1**	**1**	**1**	**1**
Stp	**1**	**1**	**1**	**1**
Stp	**1**	**1**	**1**	**1**

3 Results

Our model, based on the probability difference between different types of mutations in combination with the mutation resistance differences of the three codon positions, allowed to define a new optimization principle for a set of triplets. We examined all possible theoretical codes with fixed number of codons [4, 12, 16-23], calculated their resistance index $P\left(x_{ij}, x'_{ik}\right)^{p_i}$ and selected those which resist best to the effect of mutations. The results obtained are as follows (*Tables 2 -4*):

1. We obtained, for t = 1,2,.., 5, even 6, a unique optimal sets.
2. All of the optimal sets have a exact distribution on the code table:
- for sets with size lower than

5 synonyms have the same nucleotides in the first and second base;

- sets of 5 or 6 codons have the same second base, and in the first position there is only purine or only pyrimidine.

3. All new groups of codons obtained through the new resistance index are found among the sets of the contemporary genetic code (without those of 5 triplets). On the other hand, it is to be noted, that all the synonymous groups have, for a given size, the same optimal structure (except the cases of serine and arginine). The six codons of serine can be considered as unity of two optimal groups with size of 4 and 2. We suppose that the group of arginine is not optimal since it has in the first position either purine or pyrimidine.
4. Finally we emphasize, that the index value of optimal set of 5 triplets is less than that of 4. Otherwise, when a new codon to optimal set is added, the index increases. This observation can explain the absence of 5 triplet groups: the set is not stable and easy transforms into smaller or larger one [4-8, 18-36].

When a new resistance index is placed in the program, the results are going better. Regardless of difference between the chosen starting coons, we obtained the structure of the contemporary genetic code. The optimal theoretical configurations showed resistance indexes close to the value of the contemporary genetic code. A theoretical code obtained by the program is shown in Table 4. The sets of serine and arginine are exception again. Nevertheless, the configuration is very similar with the contemporary genetic code which is evidence in favour of the correctness of our considerations [12, 14, 18-23].

The literature also mentions "other" parameters that play the role of "evolutionary pressure" for CGC. In general, they cover all the mechanisms that encode and maintain genetic information. For example, CGC is obviously related to the translational apparatus comprised of ribosomes and mRNA, the action of which we described here schematically by the probabilities $P(x_{ij}, x'_{ik})$ and the weight matrix depending on some chemical properties of Amino acids. All these mechanisms have probably evolved in parallel with the evolution of the CGC during the early stages of life formation. Our results give reason to conclude that the relationship between the AA and the codon that encodes it is strictly specific, that the first two letters of the triplet determine the type of mRNA.

References

[1] Ahmad, M. Jung, T. and Bhuiyan A. From DNA to protein: Why genetic code context of nucleotides for DNA signal processing? *Biomedical Signal Processing and Control*, Vol. 34, 2017, pp. 44-63.

[2] Alff-Steinberger, C. The genetic code and error transmission. *Proc. Natl. Acad. Sci.U. S. A.* Vol. 64, 1969, pp. 584–591.

[3] Amirnovin, R. An analisys of the metabolic theory of the origin of the genetic code. *J Mol Evol* ., Vol. 44, 1997, pp. 473-476.

[4] Angelov K.S, et al, The impact of the extent of surgical resection on survival of gastric cancer patients, *Onco Targets Ther.* Vol. 9 2016, pp. 4687–4694.

[5] Ardell DH, Sella G On the evolution of redundancy in genetic codes. *J Mol Evol*. Vol. 53, No. 4–5, 2001, pp. 269–281

[6] Blazej P, Wnetrzak M, Mackiewicz P., The role of crossover operator in evolutionary-based approach to the problem of genetic code optimization. *Biosystems* Vol.150, 2016, pp.61–72

[7] Capt C, Passamonti M, Breton S., The human mitochondrial genome may code for more than 13 proteins. *Mitochondrial DNA A,* Vol. 27, No 5, 201, pp.3098–3101

[8] Chechetkin R., and Lobzin V. Stability of the genetic code and optimal parameters of amino acids. *Journal of Theoretical Biology*, Vol. 269, No 1, 2011, pp. 57-63.

[9] Chechetkin, R. and Lobzin, V. Local stability and evolution of the genetic code. *Journal of Theoretical Biology*, Vol. 261, 2009, pp. 643-653.

[10] Crick, F. H. Codon--anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* Vol. 19, 1966, pp. 548–555.

[11] Di Giulio, M. On the RNA world: Evidence in favor of an early ribonucleopeptide world. *J. Mol. Evol*. Vol. 45, 1997, pp. 571–578.

[12] Ding SW, Anderson BJ, Haase HR, Symons RH., New overlapping gene encoded by the cucumber mosaic-virus genome. *Virology*, Vol 198, No 2, pp. 593–601.

[13] Facchianoa A., Massimo Di Giulio. The genetic code is not an optimal code in a model taking into account both the biosynthetic relationships between amino acids and their physicochemical properties, *Journal of Theoretical Biology*, Vol.459, 2018, pp. 45-51.

[14] Freeland, S. J. and Hurst, L. D. The genetic code is one in a million. *J. Mol. Evo*l. Vol. 47, 1998, pp. 238–248.

[15] Geyer R, Mamlouk M.A., 2018. On the efficiency of the genetic code after frameshift mutations. PeerJ 6:e4825

[16] Gilis, D. Massar, S. Cerf, N. and Romman, M., Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biology*, Vol. 2, No11, 2001, pp. 1-12.

[17] Giovanni D. et al., Single-Incision Laparoscopic Nontraumatic Left Lateral Diaphragmatic Hernia Repair, *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques*, Vol. 25, No 5, 2015, pp. e166-e169

[18] Gumbel M, Fimmel E, Danielli A, Strüngmann L On models of the genetic code generated by binary dichotomic algorithms. *Biosystems* Vol. 128, 2015, pp. 9–18

[19] Haig, D. and Hurst L. D. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* Vol. 33, 1991, pp. 412–417

[20] Hervé Seligmann, Bijective codon transformations show genetic code symmetries centered on cytosine's coding properties, *Theory Biosci.* 2018, Vol. 137: 17.

[21] Ivanov, O. Ch. Genetic code and point mutations. *Studia Biophysica*, Vol. 129, 1989, pp. 63-65.

[22] Kalaidzhieva, V., Trencheva, M., Traykov, M., "Open access software for econometric studies". Collection "Intellectual Property - a Formula for Success, Creativity and Innovation", publisher "About the Letters - О писменехь", UNIBIT, 2015, pp. 290-298, ISBN: 978-619-185-160-7.

[23] Keeling, P. J. Genomics: Evolution of the Genetic Code. *Current Biology* Vol. 26, No. 18, 2016, pp. 851-853.

[24] Małgorzata Wnętrzak, Paweł Błażej, Dorota Mackiewicz and Paweł Mackiewicz, The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm, *BMC Evolutionary Biology*, 2018, pp. 192

[25] Mavrevski R., et al, "Approaches to modeling of biological experimental data with GraphPad Prism software" WSEAS TRANSACTIONS on SYSTEMS and CONTROL, vol. 13, 2018, pp. 242-247,

[26] Mavrevski R., M. Traykov, Visualization software for hydrophobic-polar protein folding model, *Scientific Visualization, accepted* , Publication 2019.

[27] Milanov P Trenchev I., Pencheva N., , Explicit description of the set of all theoretical genetic codes. *Mathematica Balkanica*, Vol. 17, No. 1-2, 2003 |pp. 157-165.

[28] Shumarova S., et al., Development of a metachronous adrenal metastasis and a brain metastasis after resection of primary lung carcinoma *Merit Research Journal of Medicine and Medical Sciences* (ISSN: 2354-323X) Vol. 4, No. 9, 2015, pp. 415-418

[29] Traykov M., S. Angelov, and Yanev, N. "A new heuristic algorithm for protein folding in the hp model" *Journal of computational biology*, vol. 23, no. 8, 2016. pp. 662.-668.

[30] Trenchev I., Milanov P., Pencheva N. The Genetic code *optimality In: Discrete Mathematics and Applicaion, Research in Mathematics and Computer Sciences, Proceedings of the Sixth International Conference, (Eds. Sl. Shtrakov, K. Denecke)*, 2002, pp. 179-190.

[31] Trenchev I. el al, Investigation Of The Relationship Between The Hydrophobicity Of An Amino Acid And Codon, Which Shall Encodes, WSEAS TRANSACTIONS on SYSTEMS and CONTROL, vol. 13, 2018, pp. 401-408

[32] Trencheva, M.,, Banking Adequacy of Banks ". *Collection "Regional Economic Cooperation and Integration of the Southeast European Countries*", *Academic Publishing House "Tsenov", Economic Addition "D. A. Tsenov* ", 2002, pp. 377-381, ISBN: 954-23-0113-8

[33] Trencheva, M. Innovative moments in credit risk analysis and management - *Eleventh International Scientific Conference of Young Scientists "The Economy of Bulgaria and the European Union: Competitiveness and Innovation" at UNWE*, Sofia, 15 December 2015, pp. 521-528, ISBN: 978-954-8590-35-8

[34] Trifonov EN (2008) Tracing life back to elements. *Phys Life Rev*., Vol. 5, No 2, 2008, pp. 121–132

[35] Wong JT Coevolution theory of the genetic code at age thirty. *BioEssays* Vol. 27, 2005, pp. 416–425

[36] Yarus M, Caporaso JG, Knight R Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem* Vol. 74, 2005, pp. 179–198