Mathematica
Balkanica

New Series Vol. 18, 2004, Fasc. 1-2

Explicit Description of the Set of All Theoretical Genetic Codes

Peter Milanov[1, 2], *Ivan Trenchev*[1], *Nevena Pencheva*[3, 4]

An aspect of the evolution of the genetic code is to minimize the number of the errors during transcription and translation. In the pertinent literature this problem is analyzed by comparing of the genetic code and the set of theoretical codes generated randomly. In this study we present an explicit description of the set of all theoretical genetic codes as a convex polytope and prove that the characteristic vectors of these codes are vertices of this polytope. Thus, the modelling, obtained by us, sheds a new light on the mathematical analysis of the optimality of the genetic code, and allows new classes of optimization problems to be formulated and investigated, including the minimization of the errors.

At the same time the natural genetic code reveals the maximum resistance towards the translation errors. The polytope description obtained by us, gives a possibility to analyze the properties of all theoretical genetic codes, to characterize their translation errors and to compare them with those of the contemporary genetic code. Our calculations confirm that the classical genetic code is closed to the optimal one, with respect to point mutations. However, the further analysis does not give much information about the mechanism of this evolution with respect to the minimization of the mutation errors.

AMS Subj. Classification: 90C10, 90C30, 90C50, 92B05

Key Words: optimization, genetic code, mutations

1. Introduction

Genetic code (GC) could be considered as a system of storage, transmission, execution and regulation of the information encoded in the genes [9, 10]. Genetics information is coded among the length of the polymeric molecule composed of only four types of monomeric units - nucleotides (adenine, guanine, cytosine and uridine). They are organized into three-letter words called codons or triplets, which are 64. Three of them have a specific function, because are utilized to start and to terminate the reading. These three codons are stop codons.

Twenty amino acids are required for the synthesis of the proteins in the life nature. The arrangement of amino acid/codon assignments results from selection to minimize the effect of errors (i.e. mistranslation and mutation on resulting proteins) [6]. This arrangement of the contemporary GC is presented in Table 1 [9].

Table 1. The contemporary genetic code.

First Nucleotide	Second Nucleotide				Third Nucleotide
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A
	Leu	Ser	Term	Trp	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	Thr	Lys	Arg	A
	Met	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

First, second and third nucleotide refer to the individual nucleotides of a triplet codon; U - uridine nucleotide; G - guanine nucleotide; C - cytosine nucleotide; A - adenine nucleotide; Term - chain terminator codon. Abbreviations of amino acids are: Gly - Glycine; Ala - alanine; Val - valine; Leu - leucine; Ile - isoleucine; Ser - serine; Thr - threonine; Cys - cysteine; Met - methionine; Asp - aspartic acid; Asn - asparagine; Glu - glutamic acid; Gln - glutamine; Arg - arginine; Lys - lysine; His - histidine; Phe - phenylalanine; Tyr - tyrosine; Trp - tryptophan; Pro - proline.

On the other hand GC has the following properties: (a) degenerate; (b) unambiguous; (c) nonoverlapping; (d) without punctuations; and (e) universal [7,14]. Since there are used all 61 codons for 20 amino acids, there must be a degeneracy in the code, i.e. same amino acids are encoded by several codons. The group of codons, which code the same amino acid is called coding or synonym set [7, 10, 12, 13]. The reading of the GC during the process of protein synthesis does not involve any overlap of the codons. Also, once the reading is started at a specific codon, there is no punctuation between codons and the message is read in a continuing sequence of the nucleotide triplet until a stop codon is reached. From the model of Watson and Crick [2], DNA has a double helical structure. Mutations are a result of the changes occurring in the nucleotide sequence. Single base changes, or the so-called point mutations, can be in the first, second or third position.

Theoretical genetic codes (TGC) are variants of the canonical code, which satisfy the properties of the GC – (a), (b), (c), (d) and (e), formulated above. Many authors describe some of the TGC, but they generate them randomly [5, 6, 10]. In other cases graph theory modelling [11] is used to propose the topological coding of proteins.

There is a hypothesis that GC is a result of the selection during the evolutionary process, where the minimization of the errors in double helical structure is the base purpose. Minimization of the effect of errors or maximization of the resistance of single mutations is a problem from the optimization theory [10, 12]. To define an optimization problem, it is necessary: (i) the objective function must be well defined (because it is the criteria); (ii) the set of characteristic matrices of the TGC must be described. According to Freeland, the number of all TGC is around 2.10^{18} [5], so it is impossible to generate all TGC. Usually, a subset of TGC is generated and compared with contemporary GC, using some criteria [1, 5, 8, 10]. Hence, the following question is important: is it possible to describe the set of all TGC explicitly? In accordance with this question the aims of this study are: (i) to investigate the possibility to describe mathematically the set of all TGC in a form of polytope on another body; and (ii) to characterize the properties of this body.

2. Results and discussion

Modelling of the set of all TGC

Let A, C, G and U are letters from the set $A = \{A, G, C, U\}$ and let M be the set of all three-letter words which could be created from A. The letters correspond to the real names of the nucleotides (for example A – adenine). The cardinality of M is 64. Also a set $a = \{a_1, a_2, \dots, a_{23}\}$ is defined, because the amino acids are 20 and the stop codons are three. Let $L = \{L_1, L_2, \dots, L_{23}\}$ be a partition of M with the following properties:

$$\cup L_i = M, \quad L_i \neq \emptyset, \quad i = 1, \dots, 23; \quad L_i \cap L_j = \emptyset, \quad i, j = 1, \dots, 23.$$

The intersection between L_i and L_j is empty, because TGC is unambiguous. The sets L_i are not empty, because any amino acid is encoded.

The set L_i is called a coding set for the element a_i. Let S be a two-dimensional matrix 16×4, whose elements are all 64 triplets (three-letter words). The matrix S is chosen for convenience. The total number of the matrices S is 64!. In this way all three-letter words of the set M could be described by two indexes in the matrix S (Table 2). Further the elements of the matrix S will be noted by S_{jt}. For example, S_{23} corresponds to GAG.

Table 2. The matrix S.

First Index	Second Index			
	1	2	3	4
1	AAA	GAA	CAA	UAA
2	AAG	GAG	CAG	UAG
3	AAC	GAC	CAC	UAC
4	AAU	GAU	CAU	UAU
5	AGA	GGA	CGA	UGA
6	AGG	GGG	CGG	UGG
7	AGC	GGC	CGC	UGC
8	AGU	GGU	CGU	UGU
9	ACA	GCA	CCA	UCA
10	ACG	GCG	CCG	UCG
11	ACC	GCC	CCC	UCC
12	ACU	GCU	CCU	UCU
13	AUA	GUA	CUA	UUA
14	AUG	GUG	CUG	UUG
15	AUC	GUC	CUC	UUC
16	AUC	GUC	CUC	UUC

For any partition L of M, the mapping $F : L \to a$ defines one of the TGC. Let x_{ijt} be a boolean variable defined as follows:

$$x_{ijt} = \begin{cases} 0 & S_{jt} \notin L_i \\ 1 & S_{jt} \in L_i \end{cases} \tag{1}$$

$$i = 1, \ldots, 23, \quad j = 1, \ldots, 16, \quad t = 1, \ldots, 4$$

(The index i corresponds to amino acids and stop codons).

We describe all synonym sets, coding one and the same animo acid and three stop codons by the following inequalities:

(2) $$\sum_{j=1}^{16}\sum_{t=1}^{4} x_{ijt} \geq 1, \qquad i = 1, \dots, 20$$

(Any amino acid could be coded, i.e. the set L_i must not be empty),

(3) $$\sum_{j=1}^{16}\sum_{t=1}^{4} x_{ijt} = 1, \qquad i = 21, \dots, 23$$

(Any stop codon is coded by one triplet),

(4) $$\sum_{i=1}^{23} x_{ijt} = 1, \qquad j = 1 \dots 16, \quad t = 1 \dots 4$$

(Every element S_{jt} (i.e. the corresponding codon) belongs to only one coding set L_i, $i = 1, \dots, 23$),

(5) $$\sum_{i=1}^{23}\sum_{j=1}^{16}\sum_{t=1}^{4} x_{ijt} = 64$$

(All elements of S are used).

The conditions (1)-(4) define a set of boolean matrices $X = \{x_{ijt}\}$, any of which is a characteristic matrix of some partition L of M. Every matrix X has 23 two dimensional layers, any of which is a characteristic vector of the set L_i, $i = 1, \dots, 23$.

Now, let us consider the condition:

(6) $$0 \leq x_{ijt} \leq 1$$

$$i = 1, \dots, 23, \quad j = 1, \dots, 16, \quad t = 1, \dots, 4.$$

¿From the theorem of Wail-Minkovski [4] it follows that (2)-(6) defines a convex polytope $P(S, a)$. $P(S, a)$ contains the set defined by (1)-(5). Now, let $y = \{y_1, y_2, \dots, y_{20}\}$ is an integer partition of 61, i.e. $\sum_{i=1}^{20} y_i = 61, \quad y_i \in \mathbf{Z}^+$.

By setting $|L_i| = y_i$, every partition y defines a class of TGC. The all classes of TGC are defined by all partitions of 61. Let us consider the following system:

$$\sum_{j=1}^{16}\sum_{t=1}^{4} x_{ijt} = y_i, \qquad i = 1,\dots,23, \tag{7}$$

$$\sum_{i=1}^{23} x_{ijt} = 1, \qquad j = 1,\dots,16, \quad t = 1,\dots,4, \tag{8}$$

$$0 \le x_{ijt} \le 1 \quad i = 1,\dots,23, \quad j = 1,\dots,16, \quad t = 1,\dots,4. \tag{9}$$

The system (7)-(9) describes a polytope $P(S, a, y)$. Further let us consider the polytope $P(S, a, y_i)$, which is a result of the fixing of the index i ($i = 1,\dots,20$) in the system (7)-(9). ¿From the theorem of Edmonds [4], it follows that the polytope $P(S, a, y_i)$ is an integer polytope and its vertices are characteristic vectors of all bases of full matroid. All bases correspond to coding set L_i of cardinality y_i, which means that $P(S, a, y_i)$ corresponds to the classes of TGC, such that $|L_i| = y_i$.

Lemma. *The vertices of polytope $P(S, a, y_i)$ are two-dimensional matrices with 0 or 1 elements.*

The proof of the Lemma follows from theorem of Edmonds.

Theorem. *The vertices of the polytope $P(S, a, y)$ are all boolean matrices, satisfying the system (7)-(9). These matrices are the characteristic matrices of TGC, which belong to the class defined by the partition $y = \{y_1, y_2, \dots, y_{23}\}$.*

Proof. To prove the theorem, we firstly verify the correctness of hypothesis that the vertices of $P(S, a, y)$ are all boolean vectors, satisfying the system (7)-(9). Now let us suppose that X is an integer matrix, which satisfies (7)-(9), but it is not a vertex, i.e. it is a convex linear combination of vectors X^r belonging to $P(S, a, y)$. That means: $X = \sum_{r=1}^{v} \alpha_r X^r$, $\sum_{r=1}^{v} \alpha_r = 1$, $\alpha_r \ge 0$. Consequently $X_i = \sum_{r=1}^{v} \alpha_r X_i^r$, $\sum_{r=1}^{v} \alpha_r = 1$, $\alpha_r \ge 0$ (layers i, $i = 1,\dots,23$), but this is impossible, because X_i is a vertex of $P(S, a, y_i)$. This follows from the Lemma.

Conversely. Let X be a vertex of $P(S, a, y)$. If we suppose that the vertex X has a noninteger coordinate, then for some i it corresponds X_i, (i^{-th} layer) which is a fractional coordinate. It is not difficult to construct integer vectors Y_j belonging to $P(S, a, y_i)$ such that: $X_i = \sum_{j=1}^{r} \alpha_j Y_j$, $\sum_{j=1}^{r} \alpha_j = 1$, $\alpha_j \ge 0$.

Then X_i is not a vertex of $P(S, a, y_i)$. For all layers $i = 1, \dots, 23$ this procedure can be done. But that is impossible, because X is a vertex of $P(S, a, y)$. ■

Freeland in 1998 points out that that among the all TGC, which are 10^{18}, an optimal code appears one in a million [5]. According to our model the number of all TGC is rather different and extremely hight (64!). Similar number 10^{84}, was suggested by Di Giulio [3]. However, there is no any data which describes the set of all TGC. Our polytope description (main theorem) gives a possibility to analyze all TGC. Moreover, if the criteria of optimality is chosen, (this means that the function of our criteria is well formulated), then we have a correctly defined optimization problem. On the other hand results obtained by us need further investigation in order to define the appropriate objective functions, that will characterize the resistance of TGC against point mutations. The solution of these problems will give us ground to characterize more deeply the properties of the contemporary genetic code.

References

[1] A. A. Arzamastsev, Nature of the optimum DNA code, *Biofizika*, **42** (1997), No 3, 611–614.

[2] F. H. C. Crick, The origin of the genetic code, *J. Mol. Biol.*, **38** (1968), 367–379.

[3] M. Di Giulio, The non-universality of the genetic code: the universal ancestor was a progenote, *J. Theor. Biol.*, **209(3)** (2001), 345–349.

[4] V. Emilichev, M. Kovaliov, M. Kruzcov, Integer points of polytopes, In: *Polytope, graphs, optimization*, Moskva, Nauka (in Russian), 1981, 106–167.

[5] S. J. Freeland, L. D. Hurst, The genetic code is one in a million, *J. Mol. Evol.*, **47** (1998), 238–248.

[6] S. J. Freeland, R. D. Knight, L. F. Landweber, L. D. Hurst, Early fixation of an optimal genetic code, *Mol.Biol.Evol.,* **17(4)** (2000), 511–518.

[7] G. Gamov, Possible mathematical relation between deoxyribonucleic acid and proteins structures, *Nature*, **173** (1954), 318–323.

[8] D. Gilis, S. Massar, N. J. Cerf, M. Rooman, Optimality of the genetic code with respect to protein stability and amino-acid frequencies, comment reviews reports deposited research interactions information refereed research, *Genome Biology*, **2(11)** (2001), 105–111.

[9] D. K. Granner, Protein synthesis and genetic code, in *Harper's Biochemistry*, R.K. Murray, D.K. Granner, P.A. Mayes and V.W. Rodwell (Eds.), twenty-fourth edition, 1996, 118–186.

[10] O. Ch. Ivanov, P. Milanov, P. Kenderov, Genetic code optimality

from mathematical and evolutionary point of view, *Compt. Rend. Acad. Sci. Bulg.*, **40** (1987), 25–32.

[11] V. A. Karasev, V. E. Stefanov, Topological nature of the genetic code. *J. Theor. Biol.*, **209(3)** (2001), 303–317.

[12] A. A. Knight, S. J. Freeland, L. F. Landweber, Selection, history and chemistry: the three faces of the genetic code, *Trends Biochem. Sci.*, **24(6)** (1999), 241–247.

[13] I. Trenchev, P. Milanov, N. Pencheva, The genetic code optimality, in *Discrete mathematics and application, research in mathematics and computer sciences*, Proc. Sixth Intern. Conf., (Eds. Sl. Shtrakov, K. Denecke), 2002, 179–190.

[14] J. Wong, Evolution of the genetic code, *Microbiol. Sci.*, **5(6)** (1988), 174–181.

Received 30.09.2003

[1] *Department of Informatics,*
Faculty of Mathematics and Natural Sciences,
South-West University, Blagoevgrad 2700,
66 Ivan Mihailov Str., Bulgaria
e-mail: peter_milanov@hotmail.com, milanov@aix.swu.bg
e-mail: wonther2000@yahoo.com

[2] *Department of Operational Research,*
Institute of Mathematics and Informatics,
Acad. G. Bonchev Str. bl. 8,
Bulgarian Academy of Sciences,
Sofia 1113, Bulgaria
e-mail: peter_milanov@hotmail.com

[3] *Laboratory of Peripheral Synapses,*
Institute of Physiology,
Acad. G. Bonchev street bl. 23,
Bulgarian Academy of Sciences,
Sofia 1113, Bulgaria
e-mail: nevena_pencheva@hotmail.com

[4] *Department of Kinesitherapy,*
Faculty of Pedagogy,
South-West University,
66 Ivan Mihailov Str.,
Blagoevgrad 2700, Bulgaria