# Genetic Code, Attributive Mappings and Stochastic Matrices

Matthew He
Division of Math, Science and Technology
Nova Southeastern University
Ft. Lauderdale, FL 33314, USA
Email: hem@nova.edu

***Abstract:*** In the last decade the field of mathematical and computational biology has expanded very rapidly. Biological research furnishes both data on and insight into the workings of biological systems. However, qualitative and quantitative modeling and simulation are still far from allowing current knowledge to be organized into a well-understood structure. In this paper we construct three primitive mappings based on three genetic attributes. We then apply the mappings and basic addition operation to the universal genetic code to generate three 8x8 matrices. These square matrices are stochastic in nature and demonstrate fractal similarity properties. Furthermore the powers of these matrices are also stochastic. They resemble the similar properties to the original stochastic matrices.

***Keywords*:** Genetic code, DNA, RNA, stochastic matrix, doubly stochastic matrix.

## 1. Introduction

The digital information that underlines biochemistry, cell biology, and development can be represented by a simple string of G's, A's, T's and C's. This string is the root data structure of an organism's biology. In a very real sense, molecular biology is all about sequences. First, it tries to reduce complex biochemical phenomena to interactions between defined sequences. The ultimate rational behind all purposeful structures and behavior of living things is embodied in the sequence of residues of nascent polypeptide chains. In a real sense it is at this level of organization that the secret of life (if there is one) is to be found. As soon as Watson and Crick proposed the double helix model of DNA in 1953, scientists began to study the problem of how a linear or helical DNA molecule could encode a linear protein molecule. Cracking the genetic code became a hot topic and attracted mathematicians, computer scientists, physicists even George Gamow (of the Big Bang Theory). The sequence of insulin was the only protein sequence available and it was scrutinized very carefully.

A mathematical view of genetic code is a map

$$\mathbf{g}: X \rightarrow Z,$$

where $\mathbf{X} = \{(x_1 x_2 x_3): x_i \in \mathbf{Y} = \{A, C, G, U\}\}$ = the set of codons and $\mathbf{Z}$ = {Ala, Arg, Asp, …, Val, UAA, UAG, UGA}= the set of amino acids and termination codon. The inheritable information is encoded by the texts from three-alphabetic words - *triplets* or *codonums* compounded on the basis of the alphabet consisted of four characters being the nitrogen bases: A (adenine), C (cytosine), G (guanine), T (thiamine). Using three-alphabetic triplets or codonums we can code 20 amino acids. There exist $4^3$ = 64 different combinations from four on three basis's. In this connection some of 20 amino acids are encoded at once by several triplets. It is called as a degeneracy of a code. The finding of conformity between triplets and amino acids (or signs of the punctuation) is customary treated as decryption of genetic code.

The RNA plays a role of "intermediary" in synthesis of proteins from amino acids of the 20 kinds pursuant to sequence of triplets in DNA-circuits. A well-known difference of RNA from DNA is the fact that the standard set of nitrogen bases of its triplets contains instead of a thiamine (T) an uracil (U), which is very similar and related with it, that is why the four-alphabetic code alphabet for RNA consists of the set A, C, G, U. Proteins are main dense component of alive organism. Each of proteins executes only own, appropriate to it function. The proteins represent themselves large polymer molecules consisting of circuits of amino acids (polypeptides), irregularly alternated. A molecule of protein often is compared to a train consisting of cars of twenty different kinds, which are spanned one another by the same way permitting to connect cars in any order. The following table gives a complete list of 64 triplets (codons).

| | Second nucleotide | | | | |
|---|---|---|---|---|---|
| | U | C | A | G | |
| | UUU Phenylalanine (Phe) | UCU Serine (Ser) | UAU Tyrosine (Tyr) | UGU Cysteine (Cys) | U |

table of genetic code" constructed by Petoukhov in his book in "Biperiodical table of genetic code and a number of protons" (2001).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CCC | CCA | CAC | CAA | ACC | ACA | AAC | AAA |
| CCU | CCG | CAU | CAG | ACU | ACG | AAU | AAG |
| CUC | CUA | CGC | CGA | AUC | AUA | AGC | AGA |
| UCC | UCA | UAC | UAA | GCC | GCA | GAC | GAA |
| CUU | CUG | CGU | CGG | AUU | AUG | AGU | AGG |
| UCU | UCG | UAU | UAG | GCU | GCG | GAU | GAG |
| UUC | UUA | UGC | UGA | GUC | GUA | GGC | GGA |
| UUU | UUG | UGU | UGG | GUU | GUG | GGU | GGG |

**Table 2 Biperiodic Table of Genetic Code**

The symmetrical structures of the genetic code were recently studied by the author (He, 2003). Both Tables 1 and 2 can be viewed as square matrices of dimension 8x8 with the elements of codons. The theory of matrices played important roles in many areas. In particular, the class of stochastic matrices reveal great data structure.

A square matrix of n x n $P=( p_{i\,j} )$ is a stochastic matrix if all entries of the matrix are nonnegative and the sum of the elements in each row or column is unity or a constant. If the sum of the elements in each row and column is unity or the same, the matrix is called doubly stochastic. The term "stochastic matrix" goes back at least to Romanovsky (1931). It plays a large role in the theory of discrete Markov chains. Stochastic matrices and doubly stochastic matrices have many remarkable properties. The properties of stochastic matrices are mainly spectral theoretic and are motivated by Markov chains. Doubly stochastic matrices have additional combinatorial structure. Here we list several important theorems (Bapat & Raghavan,1997) on stochastic matrices for its applications.

**Theorem A:** For any stochastic matrix P, $(I + P^{2} + P^{3} + \ldots P^{k-1})/k$ converge to a limit matrix Q for some matrix Q. Also, $QP = PQ = Q = Q^2$.

**Theorem B:** Any doubly stochastic matrix A can be written as a convex combination of finitely many permutation matrices; that is

$$A = a_1 P_1 + a_2 P_2 + \ldots + a_m P_m,$$

where $P_1, P_2, \ldots, P_m$ are permutation matrices and $0 = a_1, a_2, \ldots, a_m = 1$, $a_1 + a_2 + \ldots a_m = 1$. A permutation matrix can be obtained from an identity matrix by permuting its rows and columns.

**Theorem C:** If P is doubly stochastic and nonsingular, then $Q = P^{-1}$ satisfies $eQ = e$ and $Qe^t = e^t$, e is the characteristic vector and $e^t$ is its transpose.

In this paper we construct three primitive mappings based on three genetic attributes defined in Section 2. These mappings play an important role between genetic data and numerical values. We then apply the mappings associated with basic operation addition to the biperiodical table to generate three 8x8 matrices. These square matrices are stochastic and demonstrate fractal properties in terms of matrix dimensions. Furthermore the powers of the stochastic matrices (doubly stochastic) are also stochastic (doubly stochastic). These matrices reveal great properties for further exploration in genetic code.

## 2. Genetic Attribute Based Mappings

Recently Petoukhov discovered that the "elementary" four-letter alphabet of genetic code comprises three binary sub-alphabets according to three kinds of biochemical attributes (Petoukhov 2001, 2002).

The table below illustrates this attributive fact (indexes at symbols "0" and "1" are identical to number of an attribute type).

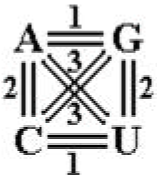
| ATTRIBUTE | **G** | **A** | **U** | **C** | |
|---|---|---|---|---|---|
| 1) Belonging to pyrimidine class (with one ring in the molecule) | $0_1$ | $0_1$ | $1_1$ | $1_1$ |  |
| 2) Amino-mutating (or special location of $NH_2$ in molecular ring) | $0_2$ | $1_2$ | $0_2$ | $1_2$ | |
| 3) Belonging to complementary pair with three hydrogen bonds | $1_3$ | $0_3$ | $0_3$ | $1_3$ | |

**Table 3 Attributes of RNA by Petoukhov**

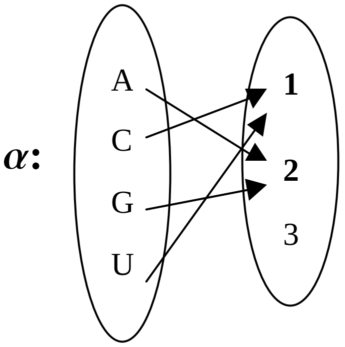
This table shows that each letter of the code alphabet has three "faces" or meanings in three binary sub-alphabets in connection with the three kinds of attributes. We'll use these attributes assign A, C, G, U values of 1, 2, and 3 for each pair of equivalence. The following lists all possible combinations of these assignments:

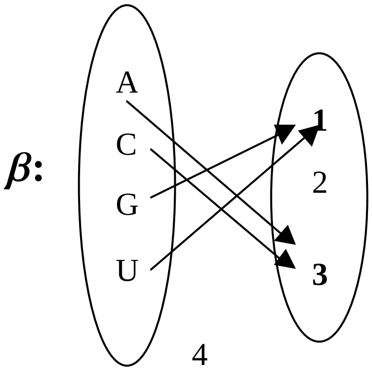
- **3.1.1 C = U = 1, A = G = 2, pyrimidines /purines ring based (1, 2)-combination**
- **3.1.2 G = U = 1, A = C = 3, amino-group based (1, 3)-combination**
- **3.1.3 A = U = 2, C = G = 3, hydrogen bonds based (2,3)- combination**

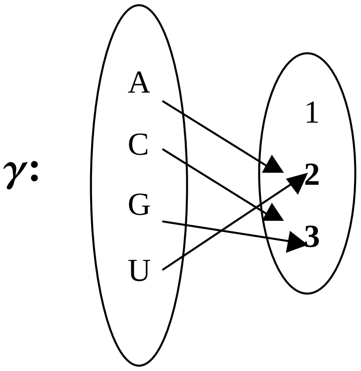
Based on these three attributes and assignments, we introduce three corresponding mappings from

$$\mathbf{Y} = \{A, C, G, U\} \text{ to } \mathbf{I} = \{1, 2, 3\}$$

as follows,





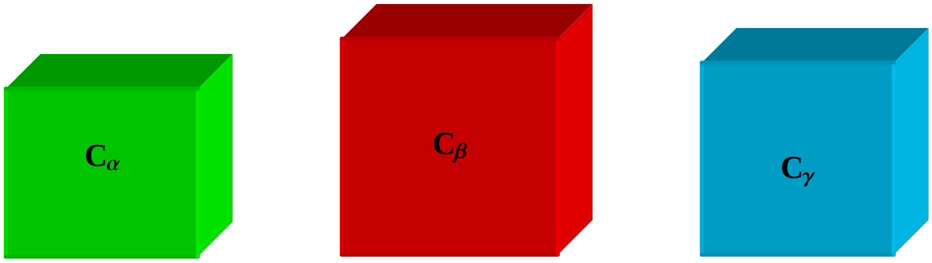


These three mappings are attribute-based functions. They map each element of $\mathbf{Y}$ to an element of $\mathbf{I}$. Applying these three mappings to each codon from the table 2, we get three tables with 64 numeric coordinates. By plotting each table in a traditional coordinate system and connecting all discrete points, we see three cubes corresponding to each table generated by three mappings $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$. These three cubes may be denoted by

$$C_\alpha = \{(1,1,1), (2,2,2), (1,2,1), (2,1,2), (2, 2, 1), (1, 1, 2), (2, 1, 1), (1, 2, 2)\},$$

$$C_\beta = \{(1,1,1), (3,3,3), (1,3,1), (3,1,3), (3, 3, 1), (1, 1, 3), (3, 1, 1), (1, 3, 3)\},$$

$$C_\gamma = \{(2,2,2), (3,3,3), (2,3,2), (3,2,3), (3, 3, 2), (2, 2, 3), (3, 2, 2), (2, 3, 3)\}.$$



In next section, we apply these mappings together with addition operation to each codon of the Table 2 to generate three sequences of matrices.

## 3. Genetic Code Based Stochastic Matrices

In Section 1 we introduced the Biperiodical table of genetic code which contains 64 codons. These 64 codons are arranged into an 8x8 matrix. In Section 2 we defined three mappings $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from $\mathbf{Y}$ ={A, C, G, U} to $\mathbf{I}$ ={1, 2, 3}. We now apply these three mappings $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively, to the table 2 with basic addition operation to generate three genetic code based matrices.

### 3.1.1. Matrix $G_8$ [1,2] (C = U = 1, A = G = 2, $\alpha$ with addition and total sum)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3 | 4 | 4 | 5 | 4 | 5 | 5 | 6 | **36** |
| 3 | 4 | 4 | 5 | 4 | 5 | 5 | 6 | **36** |
| 3 | 4 | **4** | **5** | **4** | **5** | 5 | 6 | **36** |
| 3 | 4 | **4** | **5** | **4** | **5** | 5 | 6 | **36** |
| 3 | 4 | **4** | **5** | **4** | **5** | 5 | 6 | **36** |
| 3 | 4 | **4** | **5** | **4** | **5** | 5 | 6 | **36** |
| 3 | 4 | 4 | 5 | 4 | 5 | 5 | 6 | **36** |
| 3 | 4 | 4 | 5 | 4 | 5 | 5 | 6 | **36** |
| **24** | **32** | **32** | **40** | **32** | **40** | **40** | **48** | **288** |

### 3.1.2. Matrix $G_8$ [1,3] (G = U = 1, A = C = 3, $\beta$ with addition and total sum)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | **72** |
| 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | **56** |
| 7 | 7 | **7** | **7** | **7** | **7** | 7 | 7 | **56** |
| 7 | 7 | **7** | **7** | **7** | **7** | 7 | 7 | **56** |
| 5 | 5 | **5** | **5** | **5** | **5** | 5 | 5 | **40** |
| 5 | 5 | **5** | **5** | **5** | **5** | 5 | 5 | **40** |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | **40** |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | **24** |
| **48** | **48** | **48** | **48** | **48** | **48** | **48** | **48** | **512** |

### 3.1.3 Matrix $G_8$ [2,3] (A = U = 2, C = G = 3, $\gamma$ with addition and total sum )

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | 8 | 8 | 7 | 8 | 7 | 7 | 6 | **60** |
| 8 | 9 | 7 | 8 | 7 | 8 | 6 | 7 | **60** |
| 8 | 7 | **9** | **8** | **7** | **6** | 8 | 7 | **60** |
| 8 | 7 | **7** | **6** | **9** | **8** | 8 | 7 | **60** |
| 7 | 8 | **8** | **9** | **6** | **7** | 7 | 8 | **60** |
| 7 | 8 | **6** | **7** | **8** | **9** | 7 | 8 | **60** |
| 7 | 6 | 8 | 7 | 8 | 7 | 9 | 8 | **60** |
| 6 | 7 | 7 | 8 | 7 | 8 | 8 | 9 | **60** |
| **60** | **60** | **60** | **60** | **60** | **60** | **60** | **60** | **480** |

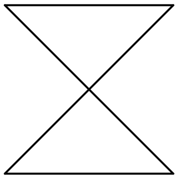
The matrix $G_8$ [2,3] was initially introduced in [3] by Petoukhov. One may notice immediately that the matrices $G_8$ [1,2], $G_8$ [1,3], $G_8$ [2,3] are stochastic matrices after normalization using the common sums. Each common sum is an eigenvalue of the matrix with the eigenvector $e^t$ = (1, 1, 1, 1, 1, 1, 1, 1). Furthermore the matrix $G_8$ [2,3] is also double stochastic matrix.

Let $S_r$, $S_c$ and $S_d$ be the common sum of the elements of rows, columns and diagonals of the matrices. We get the following table.

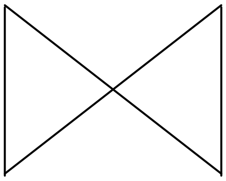
| **Matrix** | **$S_r$** | **$S_c$** | **$S_d$** | **Remark** |
|---|---|---|---|---|
| $G_8$ [1,2] | **36** | | **36** | **common row/<br>diagonal sum** |
| $G_8$ [1,3] | | **48** | **48** | **common column/<br>diagonal sum** |
| $G_8$ [2,3] | **60** | **60** | | **common row/<br>column sum** |

We may visualize these stochastic structures by using the following diagrams.

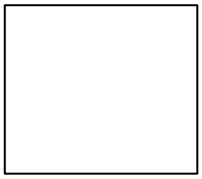
**Row sum = Diagonal sum**



**Column sum = Diagonal sum**



**Row sum = Column sum**



Next we locate the centric square matrices with dimensions of 2x2, 4x4, 6x6 and 8x8 from each matrix $G_8$ [1,2], $G_8$ [1,3], $G_8$ [2,3].

We begin with the 2x2 matrices $G_2$ [1,2], $G_2$ [1,3], $G_2$ [2,3] at the central location of the matrices $G_8$ [1,2], $G_8$ [1,3], $G_8$ [2,3], respectively. We then expand each 2x2 matrix to next level of 4x4 to generate matrices $G_4$ [1,2], $G_4$ [1,3], $G_4$ [2,3]. The next level of homogenous expansion will result in the 6x6 matrices $G_6$ [1,2], $G_6$ [1,3], $G_6$ [2,3]. One more level of expansion will give us the 8x8 matrices $G_8$ [1,2], $G_8$ [1,3], $G_8$ [2,3].

This process of expansion is illustrated below by listing all the elements of the matrices $G_2$ [2,3], $G_4$ [2,3], $G_6$ [2,3] and $G_8$ [2,3].

**$G_3$[2,3]**

| | |
|---|---|
| **6** | **9** |
| **9** | **6** |

**$G_4$[2,3]**

| | | | |
|---|---|---|---|
| 9 | 8 | 7 | 6 |
| 7 | **6** | **9** | 8 |
| 8 | **9** | **6** | 7 |
| 6 | 7 | 8 | 9 |

**$G_6$[2,3]**

| | | | | | |
|---|---|---|---|---|---|
| 9 | 7 | 8 | 7 | 8 | 6 |
| 7 | 9 | 8 | 7 | 6 | 8 |
| 7 | 7 | **6** | **9** | 8 | 8 |
| 8 | 8 | **9** | **6** | 7 | 7 |
| 8 | 6 | 7 | 8 | 9 | 7 |
| 6 | 8 | 7 | 8 | 7 | 9 |

**$G_8$[2,3]**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | 8 | 8 | 7 | 8 | 7 | 7 | 6 |
| 8 | 9 | 7 | 8 | 7 | 8 | 6 | 7 |
| 8 | 7 | 9 | 8 | 7 | 6 | 8 | 7 |
| 8 | 7 | 7 | **6** | **9** | 8 | 8 | 7 |
| 7 | 8 | 8 | **9** | **6** | 7 | 7 | 8 |
| 7 | 8 | 6 | 7 | 8 | 9 | 7 | 8 |
| 7 | 6 | 8 | 7 | 8 | 7 | 9 | 8 |
| 6 | 7 | 7 | 8 | 7 | 8 | 8 | 9 |

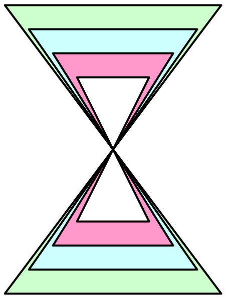
The common sums of these four matrices are 15, 30, 45 and 60. Other expanding matrices may be illustrated similarly. To unity these three types of matrices, we use the notation $G_{2n}$ [a, b], n = 1,2, 3, 4, a, b = 1, 2, 3, a < b, to represent these matrices. Specifically we arrange these matrices into a table.

| **$G_{2n}$ [a, b]** | **n = 1** | **n = 2** | **n = 3** | **n = 4** |
|---|---|---|---|---|
| **$G_{2n}$ [1,2]** | **$G_2$ [1,2]** | **$G_4$ [1,2]** | **$G_6$ [1,2]** | **$G_8$ [1,2]** |
| **$G_{2n}$ [1,3]** | **$G_2$ [1,3]** | **$G_4$ [1,3]** | **$G_6$ [1,3]** | **$G_8$ [1,3]** |
| **$G_{2n}$ [2,3]** | **$G_2$ [2,3]** | **$G_4$ [2,3]** | **$G_6$ [2,3]** | **$G_8$ [2,3]** |

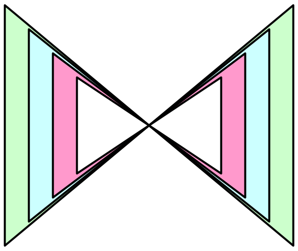
These matrices **$G_{2n}$ [a, b]**, n = 1,2, 3, 4, a, b = 1, 2, 3, a < b, are corresponding to the color coded (white for n =1, red for n = 2, blue for n = 3, green for n = 4) matrices obtained by three mappings. It is easy to verify that all these matrices are stochastic matrices and resemble the stochastic similarity. The corresponding common sums of these matrices are also listed below.

| **$G_{2n}$ [a,b]** | **n = 1** | **n = 2** | **n = 3** | **n = 4** |
|---|---|---|---|---|
| **$G_{2n}$ [1,2]** | **9** | **18** | **27** | **36** |
| **$G_{2n}$ [1,3]** | **12** | **24** | **36** | **48** |
| **$G_{2n}$ [2,3]** | **15** | **30** | **45** | **60** |

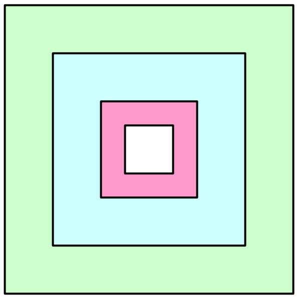
We illustrate this type of similarity by using geometrical diagrams for n = 1,2,3, and 4.



**$G_{2n}$ [1,2]**



**$G_{2n}$ [1,3]**



**$G_{2n}$ [2,3]**

Next we note that the powers (m = 1, 2, 3,…) of matrices $G^{m}_{2n}$ [a, b], n = 1,2,3,4, a, b =1,2,3, a<b, are also stochastic and the common sums of the square matrices are equal to the power of the corresponding sum of the original matrix. For example, the common sums of the matrices $G_2$ [2,3], $G_4$ [2,3], $G_6$ [2,3] and $G_8$ [2,3] are 15, 30, 45 and 60, respectively. Then the common sums of the matrices $G^2_2$ [2,3], $G^2_4$ [2,3], $G^2_6$ [2,3] and $G^2_8$ [2,3] are $15^2 = 225$, $30^2 = 900$, $45^2 = 2025$ and $60^2 = 3600$, respectively.

Our study showed a close relation between genetic code and stochastic matrix by using genetic attribute based attributive mappings. It is hoped that these relationships will help us explore the structure of genetic code. Recent advances reveal the first fact of biological sequence analysis. In biomolecular sequences (DNA, RNA, or amino acid sequences), high sequence similarity usually implies significant functional or structural similarity. However, there is not a one-to-one correspondence between sequence and structure or sequence and function. Structural or functional similarity does not necessarily imply sequence similarity. Quite distinct sequences can produce remarkably similar structures. Evolutionarily and functionally related molecular strings can differ significantly throughout much of the string and yet preserve the same three-dimensional structure(s), or the same two-dimensional substructure (s) (motifs, domains), or the same active sites, or the same or related dispersed residues (DNA or amino acid).

Life is based on a repertoire of structured and interrelated molecular building blocks that are shared and passed around. The same and related molecular structures and mechanisms show up repeatedly in the genome of a single species and a cross a very wide spectrum of divergent species. The matrices are storages of digital data. The matrices appear in various dimensions with different shapes. Stochastic matrices motivated by language of probability show up repeatedly in the nature.

## Reference

Bapat, R.B., Raghavan, T.E.S. (1997), Nonnegative Matrices and Applications, Cambridge University Press.

He, M. (2003), Symmetry in Structure of Genetic Code. *Proceedings of the Third All-Russian Interdisciplinary Scientific Conference "Ethics and the Science of Future. Unity in Diversity"*, February 12-14, Moscow.

Petoukhov, S.V. (2001), *The Bi-periodic Table of Genetic Code and Number of Protons*, Foreword of K. V. Frolov, Moscow, 258 (in Russian).

Petoukhov, S. V. (2002), Binary sub-alphabets of genetic language and problem of unification bases of biological languages, *IX International Conference "Mathematics, computer, education"*, Russia, Dubna, January 28-31, 191 (in Russian).

Romanovsky, V. (1931), Sur les zeros des matrices stocastiques, C. R. Acad. Sci. Paris 192, 266-269. [Zbl. 1 (1932) 055]

Wittmann, H.G. (1961) Ansatze zur Entschlusselung des Genetishen Codes. – *Die Naturwissenschaften*, B.48, 24, S. 55.