Available online at www.scholarsresearchlibrary.com

Scholars Research Library

Der Pharmacia Lettre, 2017, 9 [8]: 40-54
[http://scholarsresearchlibrary.com/archive.html]

ISSN 0975-5071
USA CODEN: DPLEB4

Mathematical Models for Studying the Properties of the Genetic Code

Ivan Trenchev[1], Metody Traykov[1], Miglena Trencheva[1], Madlen Stoyanova[2,3], Radoslav Stoev[4]

[1]South-West University "Neofit Rilski" Bulgaria

[2]Medical University of Sofia, Bulgaria

[3]Specialized Hospital for Pulmonary Diseases 'Sveta Sofia', Bulgaria

[4]University of Library Studies and Information Technologies, Bulgaria

***Corresponding author:** Ivan Trenchev, South-West University "Neofit Rilski" Bulgaria. E-mail: wonther2000@yahoo.com.*

ABSTRACT

The contemporary genetic code can be considered as a system for storage, processing and retrieval of the genetic information. The codes, which have the properties of the genetic code, but do not occur in nature, are called theoretical genetic codes. Their number is around 1084.

This article will make a brief retrospect of the developed mathematical models to study the properties of the genetic code. We try to find an answer of the question of why nature has chosen contemporary genetic code to the other theoretical codes.

Keywords: Solubility, Dissolution, Simvastatin, Co-crystals, Conformers

INTRODUCTION

The concepts "model" and "modeling" have been used since the 60s of the 19th century in math, physics, chemistry, biology, technology and other sciences, as their interpretation and diversity are not synonymous. In general, a model denotes a deliberate creation of artificial image of a real object or some of its properties and characteristics, while the concept of modeling generalizes activities associated with the construction and usage of the model. A substantial feature for the model is the degree of accuracy, respectively plausibility with the real object, which is highly dependent on the paradigms in a particular scientific field.

A subject of modeling in the current paper is Gene Mapping (GM). It is widely known that the knowledge in the field of molecular genetics (genomics) increases exponentially. It can be assumed that the modern age is of molecular biology and genetics, and it starts with the discovery of DNA's double helix in 1953 by Watson and Crick [1]. This revolution in biology provoked an intellectual revolution in different scientific fields and active studies, which led to the creation of the so-called molecular biotechnologies. They developed very fast and completely changed the methodology especially in sciences associated with biological subject research. Double helix sections of DNA were sequenced from different biological species – microorganisms, plants and animals. For a better interpretation of such data and for information description of such nature the purely biological approach for research was not sufficient. This led to a development of a new approach and models - mathematical. The possibility for prediction of some properties of various studied subjects expanded with them. Instances of such models in biology after the description of DNA's double helix include: Prediction of RNA's secondary structure, the protein synthesis in which the information invested in the code is translated, many problems in the biological essence of the human pathology and so on. It assisted the study of the genetic nature of some heritable diseases, inherent susceptibilities to diseases and so on, associated with mutations. Thus, new aspects in medical genetics emerged like gene therapy, and drug forms for causal treatment of genetic diseases.

Long DNA regions of various organisms were sequenced i.e., bacteria, viruses and multicellular organisms. It provoked the search of a number of questions such as: what exactly do the nucleic sequences code for and which amino acid and exactly why it? Many scientists turned to general research on the origin of the living organisms and the structure of Modern Genetic Code (MGC). This led to the creation of various evolutionary models via the application of mathematical approaches.

Another fundamental question is why of all possible Theoretical Genetic Codes (TGCs) nature "preferred" an MGC, which although with named modern, is millions of years old. What makes it so unique and does it change outside the planet Earth? Has nature reached its limit? If MGC is changed, will the organisms obtain other characteristics? Many research-workers began research in this aspect, particularly on how optimal is the MGC? Although different models of studying MGC are used based on statistic approaches, different possibilities, allocations and so on do not describe the whole set of MGCs. In general they study a fragment of it, because the number of all MGCs is about 10^{84}. So far modern technologies are unable to generate all possible genetic codes with their capabilities. That is why the theoretical description of all MGCs, as well as mathematical studies on the optimum of MGC, was challenges which provoked this research.

Mathematical Modeling of Optimal TGCs

Statistic models

Many scientists from different fields have studies the structure of the MGC. For instance, the first model of MGC was presented in 1954 by Gamow, a physicist [2,3]. It remains in literature with the designation of Gamow's "diamond code". The reason is that it presents the codon as a rhombus (diamond) comprised of four nucleotides, lined in the following way: the three vertices have random nucleotides assigned, the fourth is the nucleotide complementary to the one situated in the middle (Figure 1). For instance, nucleotides 2, 3 and 4 are random, and 1 is complementary to 2. Although possessing some inaccuracies, Gamow's code serves as a starting point for future research. It was used by Crick [4] and his subsequent works. In his following research, Gamow [5-7] made a detailed review of the different mathematical methods he could use to study the MGC as well as specified the statistic and experimental methods, the Monte Carlo methods for GC studies.

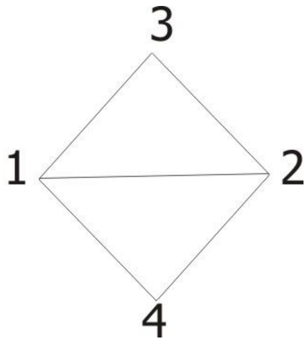

Figure 1: Schematic representation of Gamow's diamond code in 1954.

Many authors have used mathematical models for MGC research in the subsequent years [8-11]. The optimum of some criterion is studied in most research using two approaches: statistic and the alternative. The first approach uses comparison of the genetic code with random generated codes. It determines how optimal the GC on a predetermined criterion [12-18] is. The alternative approach compares MGC with the best possible theoretical codes. In general, the properties of GC are studied associated with a medium number of point mutations [19,20], and it is usually inferred that the GC is close to the optimal. Usually the first approach yields more realistic data than the second, because the alternative approach often uses linear approximations and ignores possible optimal codes. A principle subject of the optimization theories in GCs is the possible change occurrence in the genetic sequence (mutation).

In general, what the change will lead to is studied and eventual evolution of the code is inferred. The inconvenience of the two methods is that no matter what criteria of study we choose, the codes are randomly generated, and the probability of missing the optimal code is high. For instance, Haig and Hurst [21] generated 10 000 random TGCs, matching some criteria, with 64 codons allocated in 21 synonymous sets and three stop codons defined in one of the sets . These TGCs were compared to the MGC, as different scales for quantification of the AAs (amino acids) like hydrophilic properties (R_F), isoelectric properties [22], affinity to water $\log(1-R_F)/R_F$ [17], and molecular mass [23] were used.

The authors defined a medium quadratic distance in their research using the formula:

$$\sum_{xyz\in\{A,C,G,T\}^3}\left[\frac{(I)+(II)+(III)}{C}\right]$$

Where C is the number of point mutations in TGC without the stop codons being studied. (I), (II) and (III) are defined in the following manner:

$$\sum_{x^{'}\in\{A,C,G,T\}-\{x\}} D(c(xyz).c(x'yz)) \qquad (1)$$

$$\sum_{y' \in \{A,C,G,T\}-\{y\}} D(c(xyz).c(xy'z)) \qquad (2)$$

$$\sum_{z' \in \{A,C,G,T\}-\{z\}} D(c(xyz).c(xyz')) \qquad (3)$$

Where the sum is calculated only for the codons that code for AAs, meaning without the stop codons. If X and Y are AAs and X=c(*xyz*) and Y= c(*x'yz*), this means D(X,Y) = $[w(X) - w(Y)]^2$ and w(X) w(Y) are properties of an AA – hydrophilic property, molecule mass, isoelectric properties and the relation of solubility in water. They calculated four medium quadratic distances: full - MS(c); for the first position - MS_1(c); for the second - MS_2(c), and for the third - MS_3(c). The four formulas were used to calculate the full distance, for MS_1(c) - the distance the first position: the first formula for medium distance and formula (I). In the first formula *C* is substituted with C_1, the number of point mutations in the first position. The remaining couple of distances MS_2(c) and MS_3(c) were calculated in a similar fashion, except for the usage of (II) and (III), and the number of point mutations yields TGC to have better indices than MGC from with respect to simple mutations. The authors confirmed the AA's coded type highly depends on the codon's first position. Another positive outcome of the modeling is the meaning of the third position in the codon. It turns out it has influence on the mutation rate. They did not confirm convincingly in their research that the second position in the codon describes the polarity of an AA. The hypothesis, that MGC is a product of selection between very similar codes in minimization of the error effect, has not been proved. The probable cause is that Haig and Hurst did not use AA-occurrence probabilities in an average-set protein and did not define well the probabilities of mutation in the different codon positions.

In a similar manner, the optimum of the MGC was analyzed in [24]. They used an approach as the aforementioned, but they introduced an extra criterion, defining the nature of the point mutation, transition or transversion, to be exact. The authors generated 10^6 MGCs and calculated the mean quadratic distance. Freeland and Hurst studied the code in the following manner: (I) a generalization of the specific code susceptibility as a code error value; (II) definition the possible codes the MGC can alter in; (III) comparison of the values of the codes and the value of the MGC with their defined criteria. They demonstrated with their findings that only one code with a better quality than the MGC exists, presented in Table 1, and has very goods characteristics regarding point mutations. The authors' failure was that they examined the codes separately, meaning they did not compare the TGC and the MGC in the criterion. They proved in their subsequent research, in which Freeland actively took part, that the MGC is a product of code selection via the error mineralization in the decoding [25]. They studied the GC with the following formula used for defining the individual code error for each codon.

$$\Delta_i = \frac{\sum_{i=1}^{210} (w\alpha_i + \beta_i)\varepsilon_i}{\sum_{i=1}^{210} (w\alpha_i + \beta_i)}$$

j = 1..64

Where ε_i is the weight associated with the according studied AA, and can be calculated from the mutation matrix Pam $_{74-100}$. Another way of calculating ε_i is viewing it as a subtraction product of two different AAs' hydrophobia as a result of the change in the according codon. *W is* the weight, the number of transitions and transversions depends on and depending on the codon's position; α_i β_i - the number of transitions and transversions, depending on the substitution in the specific AA with "another" resulted from the "mutated" codon. For instance, the substitution of Ile in Met AUG → AUA $\alpha_{Met \rightarrow Ile}$ =1, AUG → AUY $\beta_{Met \rightarrow Ile}$ = 2. The authors confirmed the inference from the previous study that the probability of emergence of new code with better properties than the MGC is 10^{-6}. Another success was their research on the optimum of the MGC in terms of its "adaptability" to the decoding error minimization. The authors made a detailed analysis of their findings, but merely through the prism of hypotheses, not practically tested. Another drawback of Freeland's approach is that he made quite certain deductions, for instance that he "studied all possible TGSc", codon evolution could be predicted etc. (Table 1).

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

Table 1: TGC made by Freeland with better properties than the MGC

A quantitative measure was used in Freeland's and the coauthors' research in 1999 introduced by Goldman [26] to study the MGC, but the findings did not prove the hypotheses of the GC's origin of its evolution.

The hydrophilic/hydrophobic properties of the MGC were calculated in Di Giulio's research [27] and it has been proved that they are 68% better than the TGCs generated by him, using the scale defined by Gumbel et al. [28]. He used a similar approach to the aforementioned, but different, as the quadratic deviation is calculated using the formula.

$$\frac{MS_{mean} - MS_{low}}{MS_{mean} - MS_{code}}, \text{ where } MS(c) = \frac{\sum_{i.j} (X_i - X_j)^2 N_{i.j}}{\sum_{i.j} N_{i.j}},$$

Where X_i is the polar index of the i-indexed AA calculated on the hydrophilic/hydrophobic basis [29], $N_{i,j}$ is a number designating the number of codons coding the i-indexed AA of the synonymic set change and code the j-indexed AA caused by point mutations.

MS_{mean} is a mean arithmetic of MS(c) of a great number of generated TGCs; MS_{low} is the lowest value calculated from the G-function minimization, with the Lagrange method used. The G-function is defined in the following manner.

$$G(x_1, \ldots\ldots, x_{15}, \lambda) = MC(c) + \lambda\Phi \text{ Where } c \text{ is the TGC presented in в (Table 1) and } \Phi = \sum_{i=1}^{15} \mathrm{x}_i - 104.8.$$

The author came to the conclusion in this research that the MGC is 68% optimized and the hydrophilic property of the AA had probably a significant role in the evolution of the MGC. Di Giulio's approach was better, because many more TGCs were generated compared to Freeland's. We can mention the same drawback here; the author did not use the AA-emergence frequency in the averaged protein.

Haig and Hurst's quadratic distance [30] was used in Goldman's research [31] to study the MGC. He kept the power of the synonymic sets to generate the TGC as in the MGC meaning 3 AAs are coded by 6 codons, 4 AAs by 4 codons, 1 AAs by 3 codons, 9 AAs by 2 codons, 2 AAs by 1 codon and 3 stop codons. Goldman [32] came to the conclusion on the basis of his model that the main trend in the MGC's development was a simple mutation error minimization, but he did not make any conclusions about the synonymic set power.

Di Giulio [33] seriously criticized Freeland's papers and team [34-36] and did not approve the GC study approach. His thesis was that the number of TGCs was much bigger than the one generated by Freeland, and the inferred findings of a small number of generated TGCs were unreliable [37,38] with his coauthors proved in their research that the number of TGCs is around 10^{84} , and much bigger than the number of codes Freeland and team studied, meaning 10^{18} [39-42] further proved the possible TGCs are around 270 million.

On the other hand, Freeland and Di Giulio made their conclusions on a representative sample of the TGC's set, not by working on the whole set. They claimed the probability of finding a better TGC with better properties than the MGC, meeting a certain criteria, was between 10^{-6} and 10^{-9} [43]. The thing the two authors share, is the proposal that the probability cannot be established with a certainty, because it is impossible to study the whole set of TGCs (Table 2). Hence, the discussion of the explicit description of the whole set of TGCs remains open.

First position	**Second position**				**Third position**
	U	C	A	G	
U	4,8	x_5	x_8	x_{13}	U
	4,8	x_5	x_8	x_{13}	C
	x_1	x_5	Stop	Stop	A
	x_1	x_5	Stop	x_{13}	G
C	x_1	5,4	x_1	x_{15}	U
	x_1	5,4	x_1	x_{15}	C
	x_1	5,4	x_{10}	x_{15}	A
	x_1	5,4	x_{10}	x_{15}	G
A	x_2	x_6	12,5	x_5	U
	x_2	x_6	12,5	x_5	C
	x_2	x_6	13	x_{15}	A
	x_2	x_6	13	x_{15}	G
G	x_4	x_7	x_{11}	7,9	U
	x_4	x_7	x_{11}	7,9	C
	x_4	x_7	x_{12}	7,9	A
	x_4	x_7	x_{12}	7,9	G
Note: x1,..x15 are AAs diagonally written hydrophobic properties of AAs of the MGC.					

Table 2: The TGC used by Di Giulio for MGC study in 1989.

The statistic and the alternative method for MGC study found evidence for the no coincidental allocation of the codons in the synonymic sets coding AAs. For instance, it is proved that the code reduces the error effect in point mutations and in information transference coded in the DNA [44]. The authors did not study the power of the synonymic sets in the studies viewed so far and in most cases their studies ended in hypotheses.

Another interesting approach for a GC study includes the comparison of TGC with MGC in the criterion. For instance, in Gilis and team's research [45] the optimum of the MGC is studied by introduction of an extra variable in the criterion – the probability of AA in averaged protein emergence. They defined three criteria for MGC research by comparing the chemical properties of AAs coded by TGC and MGC, and found evidence that the optimum of the MGC depends on the probability of AA emergence in averaged protein as well [46]**.** They studied the way the GC's optimum depends on point mutations - transitions or transversions by constructing three functions.

$$\Phi^{FH} = \frac{1}{64}\left(\sum_{j=1}^{64}\sum_{k=1}^{64} p(c,c')g(a(c),a'(c'))\right)$$

$$\Phi^{faa} = \sum_{c=1}^{64}\frac{p(a(c))}{n(c)}\sum_{k=1}^{64} p(c,c')g(a(c),a'(c'))$$

$$\Phi^{faa} = \frac{1}{20}\sum_{c=1}^{64}\frac{1}{n(c)}\sum_{k=1}^{64} p(c,c')g(a(c),a'(c'))$$

Where *p(c,c')* is the probability of mutation emergence in the codon. The values of *p(c,c')* are nonzero if the codons c of the MGC and c' of the TGC differ only in one position, namely:

1. III position only - *1/N*;
2. I position, and in the point mutation is a transition - *1/N*;
3. I position, and in the point mutation is a transversion - *0,5/N*;
4. II position, and in the point mutation is a transition - *0,5/N*;
5. II position, and in the point mutation is a transversion - *0,1/N;*
6. In all other cases is 0.

N was calculated, in a way that that the sum of all probabilities 1, *g(a(c),a'(c'))* was a value function depending on the MGC's two codons and the studied TGC accordingly, a and a*'* are AAs coded by the two codons *c* and *c'* accordingly, p(a(c)) was the probability of AA emergence, n(c) was the number of codons of the synonymic set coding an AA a**.** Although possessing the listed advantages, Gilis did not study the power of the GC's synonymic sets, the definition of simple mutation emergence probability in a certain codon position, used by the authors of the paper, did not reflect the way the type of the point mutations influences the MGC. If the probability values were set, and did not depend on the GC, the research findings probably would have been good.

Mackay made research on a similar subject as well [47], namely a link between the number of codons coding the same AA, and the frequency of AA emergence was studied by Mackay [48]. The author examined 25 different proteins, and studied the frequency of the defined AA emergence, and found a direct link between the AA emergence frequency in nature and the number of coding codons. The author merely stated this link in his research without proving it scientifically.

Very good scientific findings were accomplished in the referred studies, but the links between the number of codons in the synonymic sets, and the probabilities of AA emergence in averaged protein are not fully explained. Most authors stated the facts without scientific of profound comments, for instance, some of them stated that the TGC set cannot be explicitly described [49-52].

Topologic and algebraic models

Besides the models for MGC property research examined so far, other models like the algebraic and the topological were used as well. They aimed at better explanation of the AA properties like the number of codons in the synonymic sets, the probabilities of AA emergence in an averaged protein etc. For instance, Suzuki and Gojobori [53] attempted in their paper to explain the link between a codon and an AA through the evolution or the natural selection method. The basis of the mathematical model was the number of codons in the synonymic set. A mutation matrix sized 61 × 61 was defined where the 3 stop codons were set through computer simulations with the number of codons in a synonymic set and an assessment of the mutation probability number of all codons in the set. The model proved that the number of codons, that codes for an AA of the GC has altered, directly linking to the AA type. The model, presented in the paper, described nucleic sequences through different stages of evolution of the codon properties in the code. The following characteristics were typical: the probabilities set for emergence in all generations in the entire examined population; each position was assumed to be independent from the other populations; the negative selection was merely represented in the model, this enabled the author to establish a direct link between the codons and the AA type. A research made by Suzuki and Gojobori represented links between the AA and the number of codons, and the authors attempted to interpret their results through the prism of evolution. Although they did not succeed in explaining in details, for instance, why the synonymic set of five codons doesn’t exist, the number of codons in the synonymic sets in the MGC is no more than 6.

The MGC property study utilised various mathematic approaches, for instance the modeling using artificial neural network has used methodology for GC study and the optimum-related problem in particular. A number of authors [54-57] used this modeling to study the MGC. Kuang, May and Taylor [58] made a matrix, through which the probability of an AA to be replaced by another AA was characterized.

Various topologic models are used to explain the MGC properties. Many researchers came to interesting findings by uncovering links between the probabilities of an AA emergence in an averaged protein and the number of codons in the synonymic set. For instance, Sneath [59] studied the MGC by utilizing various metrics for description of the probable point mutations in the 2nd, 1st and 3rd position of the codon in his paper. He came to important conclusions, namely that the probabilities of point mutations emergence vary in the 1st, 2nd and 3rd position of the codon. This gave him grounds to state his hypothesis that a minimization error is probably one of the basic driving forces of codon evolution in nature, but he never proved it. The position differentiation in the codon was a huge advantage in the developed models i.e., the various probabilities of point mutation emergence, although they did not explain the reasons of the MGC’s degeneracy and the power of synonymic sets in the MGC.

The code’s degeneracy is also a subject of active research interest [60]. interpreted the GC degeneracy via an algebraic model through Lee’s simple algebra. The authors’ hypothesis aimed at explaining the MGC’s structure comprised of 3 sets of 6 elements, 5 of 4 elements, 2 of 3 elements, 9 of 2 elements and 2 of 1 element born of Lee’s simple 64-element algebra. The authors did not explain precisely the power of the synonymic sets. Hence, the codon allocation deserves a more profound research.

The link between the AA number and the code evolution has also been studied in literature. For instance, [61] examined an interesting MCG evolutionary model, in which the code became more complex with the rising number of AAs. The presented interesting hypotheses on the codon change. For example, the possible evolution of the GCA codon coding Alanine (*Ala)*, in the

CAG codon coding Glycine (*Gly)* and so on. There is an interesting formula that calculates the probability of one AA, called primary, coding one synonymic set, changing into another AA, called secondary. It looks like this.

$$P = \sum_{x}^{n} \frac{a!}{(a-x)!x!} x \frac{b!}{(b-n+x)!(n-x)!} x \frac{(a+b-n)!n!}{(a+b)!}$$

where *a* is the total number of codons with one point mutation in the primary AA; *b* is the number of codons with more than one point mutation in the primary AA; *x* is the number of secondary AA's codons calculated from the primary AA with one point mutation; *n* is the number primary AA's codons. The calculations gave the authors grounds to state the position that one of the code's evolutionary tendencies is to minimize probable point mutation emergence without specifying the reasons for that. They stated that their findings were theoretical and can be difficult to test in practice.

MGC and TGC stability – Evolutionary aspects

The GC modeling is a subject of various methodological approaches in various knowledge fields. The findings, however, are linked to the organic world evolution in the living nature. The MGC is universal to all living organisms [62] although some differences are found in some bacteria and mitochondrial structures. The differences lie in some codons coding another AA, but in the general case it is agreeable that the MGC is universal. [63].

The idea that the MGC has evolved, was presented in an interesting way [64]. He believed the early codes coded for merely a few AAs and the number of codons has grown with the code's evolution. Many and various hypotheses have been stated to explain the possible code evolution to the MGC and some of its current properties [65-67]**.** They raised the question of the MGC's optimum in their research and whether the code is really optimal or has nature preferred it indeed because of its unique properties, namely the evolutionary abilities left inside it. Other authors interpreted the MGC emergence with the following error minimization scenario in transitions and transversions during the MGC's change [68]. To prove that hypothesis, many MGC researchers took coefficients describing the percentage of MGC optimum according to the number of changes with a different value caused by point mutations [69-71]. The authors stated in most cases that the MGC is optimal and finding a TGC with better properties is unlikely.

RESULTS AND DISCUSSION

It has been experimentally proved that point mutations occur more frequently in I and III codon position than in the II codon position. [67,58,30, 31,39]. It is an interesting scientific fact used in optimum MGC research. Many authors described the MGC properties and made interesting hypotheses. For instance, in their review [72] described the MGC's properties and the authors stated the following suggestions: (I) similar codons code for the same AA and differ merely in the third position; (II) codons that have U in the second positions code for hydrophobic AAs, and those with A in that position code for hydrophilic. They never proved these suggestions, but merely stated them, nevertheless they deserve profound research.

The authors did not compare the TGC to the MGC in some studies, but merely calculated the TGC error, for instance [70] and [71] held that the GC has developed with respect to error minimization in message passing (translation) and simple error mutations (transitions, transversions). Findings from those studies do not fully support the originally set hypotheses. Other authors also claim the number of the synonymic sets in the GC is not coincidental [68]. They examined the link between the AA and the codon type, and found a biosynthetic link i.e., a link between the chemical structure of the AA and the codon coding for it. The findings in these two articles confirmed the link between the AA type and the codon, but the authors did not comment on the reason for the link, what it was caused by.

Other authors, who worked on a similar subject, attempted to find evolutionary aspects regarding the link between the codons and monogenic structures similar to a certain AA [68,55]. For example, an AA pertaining to a specialized group containing: *pyruvate*, *aspartame* and *glutamate*, the codons have in their first position U, G, A and C. Many authors attempted to analyze these structures regarding the GC's optimum [66], but in most cases the findings were unsatisfactory. The fact that AAs with similar structures and biochemical properties are coded by similar codons explains the necessity of such a link [62,66]. The minimization error in simple mutations or code reading error is probably a main tendency [60,32].

The evolutionary aspects in the GC stability are also studied through the code degeneracy analysis. For instance, [33] presented the GC as a link between two sets (one of 64 elements, the other of 20 elements). They studied the code's degeneracy, as well as the probable mutations in the third position. That way they inferred that the third position is crucial to the code, while the first and the second positions do not differ much in their significance. An interesting finding was also studied by Crick [7], who defined the term *downstream* MGC analysis, and studied the number and properties of codons in terms of the question why similar codons code for the same AA and the importance of the codon positions. Although not fully explaining the link between the AA and the codons, the study made thorough assessment of the possible reasons for the occurrence of the link.

Another method for GC synonymic set research applied in biochemistry and molecular biology, is constructing an acceptable algorithm to study the origin of the GC and to explain the link between the AA and the synonymic set of codons coding for it [60,57,58,62].

Some authors attempted to answer the question of the differences between the codons in a synonymic set. They proposed in this regard scenarios of the early coevolution in the "ribonucleic world" i.e., that the AA evolution and the nucleic evolution is integral [12,13,64,70]. The fact that the number of tRNA is greater than the number of AAs in these studies is nevertheless neglected.

Some classic models, attempting to explain the “philosophy” of the evolutionary process, are known in literature. For instance, one such model is the “lethal model” [52]. According to the model the code error minimization or the so-called natural selection idea i.e., the natural evolution – a mutation in the code occurs and that automatically leads to a change in the protein, but the organism does not always accept the new protein and the genetic code alters with an error minimization. The lethal model hypothesis states that organisms with mutated nucleotides are in most cases unviable [65]. Created a similar model called “transport model”, because it explains the translation and it also explains that the transportation error minimization is the main factor in the GC development.

Another model, known in literature, is the so-called Wobble hypothesis. It states that the code degeneracy is a result of codon – anticodon link simplification, i.e., to minimize the number of tRNA recognizing 1 AA through evolution. All the models explained the code's evolution from different points of view, but probably the truth lies in between.

Interesting studies are made in microorganisms. Their short life enables the researchers to study mutations in different scientific aspects. An abundance of experimental data exists in which it is proved that mutations in bacteria are "beneficial" from biological perspective, because they enable them to express genes provoking the protein synthesis beneficial to the organism in the environment. That way the mutations head for the genes leading to genotype alterations and are of interest in the current research. The experiments prove that simple organisms can plan the mutations depending on the environmental pressure which aids their survival [5]. An interesting finding, which infers that the microorganisms adapt depending on the environment. The statement is disputed in some в scientific fields.

CONCLUSION

In this paper, we have provided a complete modeling of the evolution process of contemporary genetic code. We have restricted our attention only to substitution of little type of mathematical models for investigation optimality of genetic code.

ACKNOWLEDGMENTS

This work is partially supported by the project of the Bulgarian National Science Fund, entitled: Bioinformatics research: protein folding, docking and prediction of biological activity, code NSF I02/16, 12.12.14.

REFERENCES

1. Ahmad, M., Jung, L., and Bhuiyan A., From DNA to protein: Why genetic code context of nucleotides for DNA signal processing? *Biomedical Signal Processing and Control*. **2017.** 34: 44-63.
2. Alff-Steinberger, C.,The genetic code and error transmission. *Natl. Acad. Sci.* USA **1969.** 64: 584–591.
3. Amirnovin, R., An analisys of the metabolic theory of the origin of the genetic code. *J Mol Evol*, **1997.** 44: 473-476.
4. Ardell, DH., On error minimization in a sequential origin of the genetic code. *J Mol Evol.* 47, 1–13.
5. Cairns, J., Overbaugh, J. and Miller, S. *Nature* (London), **1998.** 335: 142-145.
6. Crick, FHC., The origin of the genetic code. *J Mol Biol.* **1968.** 38: 367–379.
7. Crick, FHC., et al. A speculation on the origin of protein synthesis. *Orig. Life*. **1976.** 7: 389–397.
8. Chechetkin, VR.., Lobzin, VV., Stability of the genetic code and optimal parameters of amino acids. *Journal of Theoretical Biology*, **2011.** 269(1): 57-63.
9. Chechetkin, VR., Lobzin, VV., Local stability and evolution of the genetic code. *Journal of Theoretical Biology*, **2009.** 261(4): 643-653.
10. Giulio, D., The origin of the genetic code in the ocean abysses: New comparisons confirm old observations. Journal of Theoretical Biology, **2013.** 333: 109-116.
11. Giulio, MD., The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J. Mol. Evol.* **1989.** 29: 288–293.
12. Giulio, MD., On the origin of the genetic code. *J. Theor. Biol*. **1997.** 187: 573–581.
13. Giulio, MD., On the RNA world: Evidence in favor of an early ribonucleopeptide world. *J. Mol. Evol*. **1997.** 45: 571–578.
14. Giulio, DM., and Medugno, M., The historical factor: the biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code. *J. Mol. Evol.* **1998.** 46: 615–621.

15. Giulio, MD., et al. On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. *J. Theor. Biol*. **1994.** 168: 43–51.
16. Giulio, D., The coevolution theory of the origin of the genetic code. *Physics of Life Reviews*, **2004.** 1(2): 128–137.
17. Epstein, CJ.,. Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. *Nature.* **1966.** 210:25-28.
18. Freeland, SJ., Knight RD., and Lanweber LF., Measuring adaptation within the genetic code. Tibs, **2000.** 25: 44-45.
19. Freeland, SJ., The Darwinian genetic code: An adaptation for adapting? *Gen. Prog. Evol. Mach*, **2002.** 113–127.
20. Freeland, SJ., Knigh, RD., and Landweber, LF.,. Do proteins predate DNA? Science, **1999.** 286:690–692.
21. Freeland, SJ., Hurst, LD., The genetic code is one in a million. *J. Mol. Evol*. **1998.** 47: 238–248.
22. Gamow, G., Possible mathematical relation between deoxybonucleic acid and proteins. Det Kongelige Danske Videnskabernes. *Biologiske Meddelelser*, **1954.** 22:1-13.
23. Gamow, G., Possible relation between deoxyribonucleic acid and protein structures. *Nature,* **1954.** 173: 318.
24. Gamow, G., Rich, A., and Ycas, M., The problem of information transfer from nucleic acids to proteins. *Advances in Biological and Medical Physics*, **1956.** 4: 23–68.
25. Gilis, D., et al. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biology, **2001.** 2(11): 45-57.
26. Goldman, N., Futher results on error minimizationin genetic code. *J Mol Evol*. **1993.** 37:662 -670.
27. Grantham, R., Amino acids different formula to help explain protein evolution. Science, **1974.** 185: 862-864.
28. Gumbel, M., et al. On models of the genetic code generated by binary dichotomic algorithms. *Biosystems*, **2015.** 128: 9-18.
29. Haig, D., and Hurst, LD., A quantitative measure of error minimization in the genetic code. *J. Mol Evol.* **1991.** 33: 412–417.
30. Och, I., Genetic code and point mutations. Studia Biophysica, **1989.** 129:63-65.
31. Och, I., Milanov, P., Kencderov, P. Genetic code optimality from mathematical and evolutionary point of view. Compt. Rend. *Acad. Sci. Bulg*., **1987.** 40:25-32.
32. Keeling, PJ., Genomics: Evolution of the Genetic Code. *Current Biology*, **2016.** 26(18): 851-853.
33. Knight, RD., Freeland SJ., and Landweber, F., Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci*. **1999.** 24: 241–247.
34. Knight, RD., Freeland SJ., and Landweber, F., Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci*. **1999.** 24: 241–247.
35. Kuang, L., et al. Amino acid substitution matrices from an artificial neural network model. The Ridgeway, Mill Hill NW7 1AA UK, **2001.**
36. Kuruoglu, EE., Arndt PF., The information capacity of the genetic code: Is the natural code optimal? Journal of Theoretical Biology, **2017.** 419: 227-237.
37. Lenstra, R., Evolution of the genetic code through progressive symmetry breaking. *Journal of Theoretical Biology*, **2014.** 347: 95-108.
38. Mackay, AL., Optimization of the genetic code. *Nature*., **1967.** p. 216.
39. Michael, F., and Sebastian, S., Lie superalgebras and the multiplet structure of the genetic code. I. Codon representations. *J. Math. Phys.* **2000.** 41(8): 5407–5422.

40. Nemzer, LR., A binary representation of the genetic code. *Biosystems, Article in Press,* **2017.**
41. Nitta, I., et al. Reconstitution of peptide bond formation with *Escherichia coli* 23 Sribosomal RNA domains. Science, **1998.** 281: 666–669.
42. Orengo, CA., et al. CATH - A hierarchic classification of protein domain structures. *Structure,* **1997.** 5(8): 1093-1108.
43. Pelc, S.R,. Correlation between coding-Triplets and amino acids. *Nature*, **1965.** 207:597-599.
44. Perlwitz, M., et al. Pattern analysis of the genetic code. *Adv. in Appl. Math.* **1988.** 9(1): 7–21.
45. Rodin, A.S., Rodin, SN., The universal genetic code and non-canonical variants. Brenner's Encyclopedia of Genetics (Second Edition), **2013.** 263-264.
46. Ronneberg, T.A., Landweber, LF., and Freeland, SJ., Testing a biosynthetic theory of the genetic code: Fact or artifact? **2000.** 97:25, 13690 -13695.
47. Salemne, FR., Miller, MD., and Jordan, JR., Structural convergence during protein evolution. *Proc Natl Acad Sci USA,* **1977.** 74:2820-2824.
48. Sander, C., and Schneider, R., Database of homology-Derived protein structures and the structural meaning of sequence alignment. *Proteins-Structure Function and Genetics,* **1991.** 9(1): 56-68.
49. Sawyerr, BA., et al. Real-coded genetic algorithm with uniform random local search. Applied Mathematics and Computation, **2014.** 228: 589-597.
50. Sciarrino, A., and Sorba, P., Codon–anticodon interaction and the genetic code evolution. *Biosystems.* 111, 3, 175-180.
51. Sciarrino A., Sorba P., (2013) Codon–anticodon interaction and the genetic code evolution. *Biosystems.* **2013.** 111(3): 175-180.
52. Sneath, PH., Relations between chemical structure and biological activity in peptides. *J Theor Biol*. **1966.** 12(2):157-95.
53. Sonneborn, TM., Degeneracy of the genetic code: extent, nature and the genetic implications. In *Evolving Genes and Proteins.* Edited by Bryson V, Vogel HJ. New York: Academic Press, USA. **1965.** 377-397.
54. Sonneborn, TM., Degeneracy of the genetic code: extent, nature and the genetic implications. In *Evolving Genes and Proteins.* Edited by Bryson V, Vogel HJ. New York: Academic Press, USA. **1965.** 377-397.
55. Speyer, JF., et al. Synthetic polynucleotides and the amino acid code. *Cold Spring Harbor Symp Quant Biol*, 28:559-567.
56. Suzuki, Y., and Gojobori, T., A method for detcting positive selection at single amino acid sites. *Mol. Biol. Evol.* **1999.** 16(10) 1315-1328.
57. Szathmáry, E., The emergence, maintenance, and transitions of the earliest evolutionary units. Oxf. Surv. *Evol. Biol.* **1989.** 6: 169–205.
58. Szathmáry, E., Useful coding before translation: The coding coenzymes handle hypothesis for the origin of the genetic code, in Evolution: from Cosmogenesis to Biogenesis, **1990.**77–83.
59. Szathmáry, E., Coding coenzyme handles: A hypothesis for the origin of the genetic code. Proc. Natl. Acad. Sci. USA. **1993.** 90: 9916–9920.
60. Szathmáry, E., Coding coenzyme handles and the origin of the genetic code. In: From simplicity to complexity in chemistry –and Beyond. Part I.(Müller, A. et al., eds), **1996.** 33–41.
61. Szathmáry, E., and Maynard Smith, J., The major evolutionary transitions. Nature, **1995.** 374: 227–232.
62. Szathmáry, E., and Maynard Smith, J., From replicators to reproducers: the first major transitions leading to life. *J. Theor. Biol.* **1997.** 187: 555–571.

63. Taylor, WR., Protein structure comparison using iterated double dynamic programming. *Protein Science,* **1999.** 8(3): 654-665.
64. Tlusty, T., A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes. Physics of Life Reviews, **2010.** 7(3): 362-376.
65. Vogelm, G., Tracking the history of the genetic code. *Science*, **1998.** 281:329-331.
66. White, HB., Co-enzymes as fossils of an earlier metabolic stage. *J. Mol. Evol*. **1976.** 7:101–104.
67. Woese CR., The genetic code: The molecular basis for genetic expression. New York: Harper and Row, USA.
68. Woese, CR., et al. On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symposium on Quantitative Biology, **1966.** 31: 723-736.
69. Woese, CR., On the evolution of the genetic code. *Proc Natl Acad Sci USA*, **1965.** 54:1546-1552.
70. Wong, JTF., Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc Natl Acad Sci USA*, **1980.** 77:1083-1086.
71. Wong, JT., A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci*. USA. **1975.** 72: 1909–1912.
72. Zhang, Z., Jun, Yu., Does the genetic code have a eukaryotic origin? *Genomics, Proteomics and Bioinformatics,* **2013.** 11(1): 41-55.