On the origin of the mitochondrial genetic code:

Towards a unified mathematical framework for the management of genetic information

Authors:

Diego Luis Gonzalez [1, 2]

Simone Giannerini [2]

Rodolfo Rosa [2,1]

Affiliations:

(1) CNR Istituto IMM, v. P. Gobetti 101, I-40129, Bologna

(2) Dipartimento di Scienze Statistiche, Università di Bologna, v. delle Belle Arti 41, I-40126 Bologna

Summary:

The origin of the genetic code represents one of the most challenging problems in molecular evolution. The genetic code is an important universal feature of extant organisms and indicates a common ancestry of different forms of life on earth. Known variants of the genetic code can be mainly divided in mitochondrial and nuclear classes. Here we provide a new insight on the origin of the mitochondrial genetic code: we found that its degeneracy distribution can be explained by using a mathematical approach recently developed for the description of the Euplotes nuclear variant of the genetic code. The results point to a primeval mitochondrial genetic code composed of four base codons, which we call *tesserae*, that, among other features, exhibit outstanding error detection capabilities. The theoretical description suggests also a formulation of a plausible biological theory about the origin of protein coding. Such theory is based on the symmetry properties of hypothetical primeval chemical adaptors between nucleic acids and amino acids (ancient tRNA's).

Our paper provides a unified mathematical framework for different hypotheses on the origin of genetic coding. Also, it contributes to revisit our present view about the evolutionary steps that led to extant genetic codes by giving a new first-principles perspective on the difficult problem of the origin of the genetic code, and consequently, on the origin of life on earth.

1) Introduction:

If all present forms of life descend from a common ancestor, the characteristics of such ancestor need to be searched among universally shared traits of extant organisms, non-universal traits being the consequence of accumulated divergence through evolutionary times. One of the most remarkable of these traits is the protein synthesis apparatus. This apparatus is responsible of the translation into amino acids of the nucleotide information contained in coding mRNA sequences according to the genetic code translation table. Unfortunately, since biochemical pathways do not fossilize, we do not have direct access to the information on ancestral biological steps that led to the present structures. As these previous biological steps are of primary importance for explaining the origin of life on earth, it may seem that the ultimate causes of such origin be hidden by such absence of evidence. Nevertheless, a similar problem which arises in Cosmology, when it comes to understand the origin of matter in our known universe, has been successfully tackled. The task

might appear an impossible one since we would need to go back to the beginning of time and we cannot reproduce the original Big Bang. However, with the discovery of sky background radiation, different theoretical hypotheses about the origin and early evolution of our universe became verifiable. Different theories can be tested according to their predictions, for example, regarding the frequency content and the spatial distribution of the relic radiation produced at the time of the Big Bang. We cannot reproduce experimentally the origin of the universe but this does not prevent us from obtaining quantitative information about how the different origin scenarios may have shaped the background radiation.

In the context of the origin of life, the challenge is to understand how different hypotheses can lead to different features of some biochemical analogous of the background radiation and of its relics. In order to accomplish this task we study one of the most universal, and ancient traits of extant life: the genetic code. We will show how the mathematical structure of the genetic code can be seen as a good candidate for understanding the origin of protein coding. We apply the cosmology analogy and we study the organization of the genetic code accordingly: hypotheses on origins are verified by comparing theoretical predictions against the empirical evidence of the present organization. We cannot reproduce the origin of life but we can propose verifiable theories. We might even hope to find "mathematical relics" that go back to the "Big Bang" of life. Perhaps it is not a simple coincidence that the theoretical physicist George Gamow, which first proposed the Big Bang theory for the origin of our universe, was also the first to propose a mathematical organization (turned out to be wrong) for the coding of amino acids along the double helix of DNA, the so called Gamow's diamond code (Gamow, 1954).

The main aim of this paper is to contribute to share some light on the emergence of the genetic code from a theoretical "first principles" point of view. The mathematical methods used pertain mainly to the fields of number theory and discrete group theory. Some key properties have been developed specifically for the present paper and represent a further advance in the theory that allowed the study of the nuclear genetic code (Gonzalez 2008, Gonzalez et al 2009). Within this framework, we are able to build a model that describes completely the degeneracy distribution of the mitochondrial genetic code. At the same time we develop a model of a primeval mitochondrial genetic code and show that our predictions can be reconciled with different former hypotheses about the origin of a simpler code in the form of doublets. In this context, we uncover several hidden symmetries of the present mitochondrial genetic code which are inherited form this primeval code. Further results in this context are remarkable and contribute with a unified view about a possible 2-doublet (*tessera)* origin of both the mitochondrial and the nuclear genetic codes. We present here for the first time these new hypotheses.

In section 2 we present an introduction to the theoretical model which is based on the so-called *non-power binary representations of integer numbers*; the application of these representations for describing the properties of the *Euplotes* nuclear genetic code is can be found in (Gonzalez, 2004-2008). In section 3, the approach is extended to the description of the mitochondrial genetic code. In section 4, we explore the possible biological foundations of the mathematical properties of the model of the mitochondrial genetic code. Finally, in section 5, we discuss briefly the possible consequences of the present approach for our knowledge about the origin of the genetic code and the protein coding apparatus.

2) A mathematical model of the nuclear genetic code

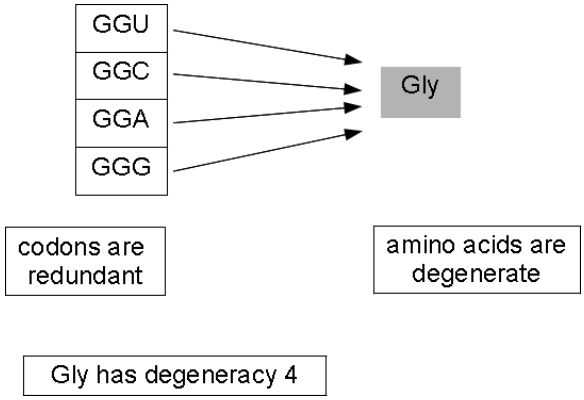
Quoting from Knight and Landweber (Knight and Landweber, 2000) :

> " [...] In the absence of evidence, many of the most interesting questions about the genetic code have fallen into a twilight zone of speculation and controversy. Although it

is generally accepted that the modern code evolved from a simpler form, there has not been consensus about when the initial code evolved or what is was like, how and when particular amino acids were added, how and when the modern tRNA/synthetase system arose, or the processes by which the code could have expanded"

A few theories for the origin of the present genetic code have been proposed. These fall in three main categories, i.e., the frozen accident (Crick, 1968), the stereo chemical one (Woese, 1965), and the co-evolution hypotesis (Wong, 1975). The frozen accident is mainly based upon the assumption that the genetic code originated by chance; successive evolution of primordial forms of life arrived at a point where any variation in the genetic code would have been deleterious; thus, the genetic code "freezes" in such form. The stereo chemical origin assumes that some kind of chemical interaction was responsible for the attachment of specific amino acids to a corresponding anti codon. With evolution, at least part of this stereo chemical specificity was lost favouring advantageous substitutions. Finally, the co-evolution theory assumes that the assignation of codons to amino-acids has followed a path that mimics the evolution of the amino acids synthetic pathways.

Box 1: the genetic code is a non bijective mapping between 64 codons and 20 amino-acids plus the stop signal. Since different codons may represent the same amino acid such codons are said *redundant* or, following a linguistic analogy, *synonymous*. On the side of amino-acids, we say that a given amino-acid is *degenerate* if it is represented by more than one codon.



In the above example the amino-acid Gly is coded by 4 codons of the kind GGN, where N can be one of the four nucleotides U,C,A,G. Thus, the four codons are redundant and the amino acid Gly has degeneracy 4. The term degeneracy originates in particle physics and refers to the existence of different quantum states having the same energy.

The description of the degeneracy of the genetic code requires the distinction of two different aspects of its organization: the first refers to the specific assignation of amino acids to codons; the

second aspect regards the global properties of the degeneracy. We know from the wobble rules (Crick, 1966), that the degeneracy of amino acids inside quartets (groups of four codons sharing the first two letters) should be 1,2,3 or 4, but it is difficult to imagine how a local random assignation of these degeneracy numbers to the 20 amino acids inside the 16 quartets of the genetic code could produce a global organization that follows some symmetry principles. However, the genetic code shows global regularities that have been partially observed early. In fact, the theoretical physicist Rumer (Rumer, 1966) noted that the standard genetic code can be divided into two halves, one containing only amino acids with degeneracy four, and the other containing amino acids with degeneracy one, two or three. Moreover, a global transformation of the bases, called Rumer's transformation, (i.e., the transformation that exchange bases maintaining their amino-keto character, U,C,A,G, ↔ G,A,C,U), changes for sure the degeneracy class of an amino acid (4↔1,2,3). Is Rumer's classification only a curious product of chance? Or, rather, is it a key property related to a deeper hidden organization of genetic information?
In the following we will show that the answer to this question can be provided by using and extending the same mathematical approach that describes exactly the degeneracy distribution of the nuclear genetic code. In particular, we provide a unified framework for both the nuclear and the mitochondrial code and show that the mathematical representation of the nuclear code can be obtained from a symmetry break in the representation of the mitochondrial code. We start from the description of the Euplotes nuclear genetic code as reported in (Gonzalez 2004, 2008, Gonzalez et al., 2009, 2012).

Non power number representation systems

As we have already pointed out, the genetic code is a non bijective mapping between two sets of different cardinality. This implies that amino-acids are degenerate and the number of amino acids that share a given degeneracy determines the degeneracy distribution. The degeneracy is explained in Box 1 whereas In Table 1 we show this distribution for the *Euplotes* nuclear genetic code inside quartets; a quartet is a group of four codons that share the first two letters. Hence, the degeneracy of an amino acid ranges from 1 to 4. Notice that in the standard version of the nuclear genetic code there are 3 amino-acids with degeneracy 6. In the Euplotes version these contribute with three units to both, degeneracy 2, and degeneracy 4, cells.

Degeneracy	# of amino acids
1	2
2	12
3	2
4	8

Table 1*: Degeneracy distribution of the Euplotes Nuclear Genetic Code. Degeneracy is considered inside quartets, that is, the groups of four codons sharing the two first bases.*

Now, can we describe the distribution of Table 1 by means of a mathematical model? The answer is positive and takes advantage of a particular kind of number representation system that we describe in the following.
Usual number representation systems use the powers of a number, the base *b*, for implementing an additive decomposition of the represented number. Our decimal representation system is a positional system of this kind that uses the powers of $b = 10$. In fact, the number 347, means that we have: $7 \times 10^0 + 4 \times 10^1 + 3 \times 10^2 = 347$. Power representation systems are univocal, that is, any

number has only one representation and any representation corresponds to only one number. As the genetic code is degenerate, we need a redundant system where a given number can have more than one possible representation. Redundant representation systems can be obtained from power representation systems in two ways: either by allowing digits values to span beyond the range {0, n-1}, or replacing the powers of the base (the positional weights in the additive decomposition) by a slowly growing series. Now, it is possible to show that the first approach cannot reproduce the degeneracy distribution of Table 1. On the contrary, there is a unique solution to the problem by using non-power representations. For a detailed description of such approach see (Gonzalez-2004, 2008). As for non-power representations see (Wolfram, 2002, Zeckendorf, 1972). In the following, we discuss those aspects of the representation that are needed for the extension of the model to the mitochondrial genetic code.

3) The Non Power Model of the Mitochondrial Genetic Code

The non power representation system that describes the degeneracy of Table 1 is a binary system of 6 digits where the first 6 powers of two (32, 16, 8, 4, 2, 1) have been replaced by the sequence (8, 7, 4, 2, 1, 1). The complete representation can be found in (Gonzalez, 2004, 2008).

On the same ground, we develop a non-power representation model for the vertebrate mitochondrial genetic code. The main difference with the nuclear code is that in the mitochondrial code there are no degeneracy-1 amino acids, that is, there are no univocal assignations of amino acids to codons. In fact, we see in Table 2 that the amino acids Met and Trp, that have degeneracy 1 in the Euplotes nuclear code, have degeneracy 2 in the mitochondrial code, i.e., codon AUA is assigned to Met and codon UGA to Trp. In the Euplotes nuclear version such codons are assigned respectively to Ile and Cys. Another difference is that a group of two stop signals is assigned to the codons AGU and AGC (assigned to Arg in the nuclear version).

Apparently, the two codes have only minor differences; however, it is well known that they are associated to very different biological structures. This might be related to the fact that both the ribosomes and the number of tRNAs in the two codes are very different. Remarkably, the mathematical approach is able to explain this difference. In fact, we will show that the mathematical representation of the mitochondrial genetic code is utterly different from that of the nuclear code. Most importantly, the representation of the nuclear code can be obtained from a symmetry break in the representation of the mitochondrial code.

The most striking difference between the two codes lies in their degeneracy. In fact, as shown in Table 3, in the mitochondrial world amino-acids can be coded either by 2 or 4 codons. Hence, we have 16 amino-acids that have degeneracy 2 and 8 amino-acids with degeneracy 4. What are the implications of such disparity at the mathematical level? First of all, the number of represented objects is always 24 so that in both representations the sum of the six non-power weights equals 23. Also, the absence of amino-acids with degeneracy 1 implies the existence of a 0-valued weight. This is indeed like a binary label which identifies two otherwise identical halves of the overall representation.

With these premises in mind we found that the set of positional weights that describe exactly the global degeneracy of the mitochondrial genetic code are: (8, 8, 4, 2, 1, 0). In fact, we can see in Table 4 that the degeneracy distribution of the non-power representation coincides with that of the mitochondrial genetic code (Table 3). Moreover, as in the nuclear code, this set represents a unique solution modulo trivial transformations

In Table 5, the complete non-power representation based on the positional weights (8,8,4,2,1,0) is

presented.
By comparing the symmetries generated by the non-power representations for the two codes we can obtain first principles information about their structure. Moreover, some additional hypotheses about the origin of genetic coding and the genealogy relation between the two class of codes can be proposed. A first symmetry found in the mathematical representation of the nuclear genetic code is the palindromic symmetry. In brief, every represented number *n* has a palindromic counterpart (*23-n*) with the same degeneracy. On the side of binary strings, if two strings are palindromic their binary representation is the complement-to-one one of each other. In order to check that the complement to one corresponds indeed to a conservation of the degeneracy, it suffices to observe that the complement of the first two digits associated to the weights (8,8) produces a different object with the same degeneracy. Moreover, the degeneracy can be inferred from the parity of these two digits. Digits 0,1 and 1,0, correspond to degeneracy four, and digits 0,0 and 1,1, correspond to degeneracy two.

	U	**C**	**A**	**G**	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	**U**
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	**C**
	UUA Leu	UCA Ser	UAA Stop	UGA Trp	**A**
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	**G**
C	CUU Leu	CCU Pro	CAU His	CGU Arg	**U**
	CUC Leu	CCC Pro	CAC His	CGC Arg	**C**
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	**A**
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	**G**
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	**U**
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	**C**
	AUA Met	ACA Thr	AAA Lys	AGA Stop	**A**
	AUG Met	ACG Thr	AAG Lys	AGG Stop	**G**
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	**U**
	GUC Val	GCC Ala	GAC Asp	GGC Gly	**C**
	GUA Val	GCA Ala	GAA Glu	GGA Gly	**A**
	GUG Val	GCG Ala	GAG Glu	GGG Gly	**G**

Table 2: *Vertebrate version of the mitochondrial genetic code. In light-green are evidenced the differences with the Euplotes nuclear version of the genetic code.*

Degeneracy	# of amino acids
2	16
4	8

Table 3*: Degeneracy distribution shared by the vertebrate, yeast, and ascidian mitochondrial genetic codes. The invertebrate, and trematode versions are also described by the same degeneracy distribution if we consider the quartet AGx assigned to Ser, as composed by two degeneracy-two doublets, i.e., ACY, ACR (with Y pyrimidine and R purine).*

Degeneracy	# of whole numbers
2	16
4	8

Table 4: *Degeneracy distribution of the non-power representation.*

Hence, In the case of the mitochondrial code, the degeneracy is explicitly determined by the two first digits of the non-power representation. Note that for the nuclear code this can be only partially inferred from these two digits, and this is a first indication of a symmetry break between the two representations.

Represented number	***Length 6 binary strings***																													
	8	**8**	**4**	**2**	**1**	**0**		**8**	**8**	**4**	**2**	**1**	**0**		**8**	**8**	**4**	**2**	**1**	**0**		**8**	**8**	**4**	**2**	**1**	**0**
0	0	0	0	0	0	0	R	0	0	0	0	0	1																
1	0	0	0	0	1	0	Y	0	0	0	0	1	1																
2	0	0	0	1	0	0	Y	0	0	0	1	0	1																
3	0	0	0	1	1	0	R	0	0	0	1	1	1																
4	0	0	1	0	0	0	Y	0	0	1	0	0	1																
5	0	0	1	0	1	0	R	0	0	1	0	1	1																
6	0	0	1	1	0	0	R	0	0	1	1	0	1																
7	0	0	1	1	1	0	Y	0	0	1	1	1	1																
8	1	0	0	0	0	0	R	1	0	0	0	0	1		0	1	0	0	0	0	Y	0	1	0	0	0	1
9	1	0	0	0	1	0	Y	1	0	0	0	1	1		0	1	0	0	1	0	R	0	1	0	0	1	1
10	1	0	0	1	0	0	Y	1	0	0	1	0	1		0	1	0	1	0	0	R	0	1	0	1	0	1
11	1	0	0	1	1	0	R	1	0	0	1	1	1		0	1	0	1	1	0	Y	0	1	0	1	1	1
12	1	0	1	0	0	0	Y	1	0	1	0	0	1		0	1	1	0	0	0	R	0	1	1	0	0	1
13	1	0	1	0	1	0	R	1	0	1	0	1	1		0	1	1	0	1	0	Y	0	1	1	0	1	1
14	1	0	1	1	0	0	R	1	0	1	1	0	1		0	1	1	1	0	0	Y	0	1	1	1	0	1
15	1	0	1	1	1	0	Y	1	0	1	1	1	1		0	1	1	1	1	0	R	0	1	1	1	1	1
16	1	1	0	0	0	0	Y	1	1	0	0	0	1																
17	1	1	0	0	1	0	R	1	1	0	0	1	1																
18	1	1	0	1	0	0	R	1	1	0	1	0	1																
19	1	1	0	1	1	0	Y	1	1	0	1	1	1																
20	1	1	1	0	0	0	R	1	1	1	0	0	1																
21	1	1	1	0	1	0	Y	1	1	1	0	1	1																
22	1	1	1	1	0	0	Y	1	1	1	1	0	1																
23	1	1	1	1	1	0	R	1	1	1	1	1	1																

Table 5: *Complete non-power representation of integer numbers from 0 to 23 associated to the set of weights (8, 8, 4, 2, 1, 0); The colours indicate the parity of the string (gray = even; white = odd);*

DINUCLEOTIDES			POWER REPRESENTATION				
BASES: A,U,C,G			BINARY DIGITS: 0,1				
FIRST	SECOND		#	8	4	2	1
U	U		**0**	0	0	0	0
U	C		**1**	0	0	0	1
U	A		**2**	0	0	1	0
U	G		**3**	0	0	1	1
C	U		**4**	0	1	0	0
C	C		**5**	0	1	0	1
C	A		**6**	0	1	1	0
C	G		**7**	0	1	1	1
A	U		**8**	1	0	0	0
A	C		**9**	1	0	0	1
A	A		**10**	1	0	1	0
A	G		**11**	1	0	1	1
G	U		**12**	1	1	0	0
G	C		**13**	1	1	0	1
G	A		**14**	1	1	1	0
G	G		**15**	1	1	1	1

Table 6: *On the right, the standard binary power representation of the 16 whole numbers (0 — 15), on the left the 16 different 2-letter words formed from an alphabet of 4 letters (U,C,A,G).*

Second, the weights of both representations contain the first 4 powers of two, i.e., (8,x,4,2,1,x), and (x,8,4,2,1,x). The weights denoted by "x" are responsible for the degeneracy of both representations (we call them the redundancy weights). In Table 6 we show the 16 whole numbers 0 – 15 represented by the sub-set (8,4,2,1). Notice that such representation is standard and univocal, hence, we can hypothesize that these 16 whole numbers represent univocally the 16 di-nucleotides (see Table 6, left). Now, if we add the redundancy weight 8, to this basic binary subset, that is, we consider the 5 weights: 8,8,4,2,1, we obtain the representation shown in Table 7. The effect of the redundancy weight 8 is to shift the represented numbers by 8 unities. Because of this, some numbers are represented by two binary strings. Thus, we have created the following degeneracy distribution.

Degeneracy	# of whole numbers
1	16
2	8

Clearly, this distribution together with the 32 binary strings represented in Table 7, form exactly one-half of the vertebrate mitochondrial genetic code. This is equivalent to an appropriate

superposition of two independent dinucleotides.

Represented number												
		8	**8**	**4**	**2**	**1**		**8**	**8**	**4**	**2**	**1**
0		**0**	**0**	**0**	**0**	**0**						
1		**0**	**0**	**0**	**0**	**1**						
2		**0**	**0**	**0**	**1**	**0**						
3		**0**	**0**	**0**	**1**	**1**						
4		**0**	**0**	**1**	**0**	**0**						
5		**0**	**0**	**1**	**0**	**1**						
6		**0**	**0**	**1**	**1**	**0**						
7		**0**	**0**	**1**	**1**	**1**						
8		**1**	**0**	**0**	**0**	**0**		**0**	**1**	**0**	**0**	**0**
9		**1**	**0**	**0**	**0**	**1**		**0**	**1**	**0**	**0**	**1**
10		**1**	**0**	**0**	**1**	**0**		**0**	**1**	**0**	**1**	**0**
11		**1**	**0**	**0**	**1**	**1**		**0**	**1**	**0**	**1**	**1**
12		**1**	**0**	**1**	**0**	**0**		**0**	**1**	**1**	**0**	**0**
13		**1**	**0**	**1**	**0**	**1**		**0**	**1**	**1**	**0**	**1**
14		**1**	**0**	**1**	**1**	**0**		**0**	**1**	**1**	**1**	**0**
15		**1**	**0**	**1**	**1**	**1**		**0**	**1**	**1**	**1**	**1**
16								**1**	**1**	**0**	**0**	**0**
17								**1**	**1**	**0**	**0**	**1**
18								**1**	**1**	**0**	**1**	**0**
19								**1**	**1**	**0**	**1**	**1**
20								**1**	**1**	**1**	**0**	**0**
21								**1**	**1**	**1**	**0**	**1**
22								**1**	**1**	**1**	**1**	**0**
23								**1**	**1**	**1**	**1**	**1**

Table 7: *One half of the vertebrate mitochondrial code degeneracy obtained from the non-power representation defined by the 5 bases: 8,8,4,2,1.*

Now, if we complete the representation with the weight 0, we obtain the complete degeneracy distribution of the vertebrate mitochondrial genetic code (see Table 5). The weight 0, does not change the value of the number represented as it simply duplicates the scheme obtained with the 5 first weights shown in table 7. The non-power representation defined by the set 8,8,4,2,1,0, represents the unique model that reproduces exactly the degeneracy of the vertebrate version of the mitochondrial genetic code.

Type	Non-power weights
mitochondrial	8,**8**,4,2,1,**0**
nuclear Euplotes	8,**7**,4,2,1,**1**

Hence, the two representations differ only in the redundancy weights as one unit on the weight 8 is moved to the weight 0 so that 8 + 0 becomes 7 + 1. This apparently simple change is responsible for all the differences between the two codes. Remarkably, the representation of the nuclear code can be

seen as a break in the symmetric and simpler structure of the mitochondrial code. This poses important fundamental questions about their origin and their relationships.

So far, we have described the degeneracy distribution of the vertebrate mitochondrial genetic code by means of a non-power integer number representation. Can we obtain more information about the mitochondrial code by pursuing this approach? The most important suggestion of such description is that the mitochondrial genetic code is composed by couples of di-nucleotides (doublets). In fact, the origin of the genetic code has been hypothesized as evolving from a simpler code with only 1 coding nucleotide, (Crick, 1968) to an intermediate version of two nucleotides (Jukes, 1973), for arriving at the present version of 3 nucleotides per codon (Crick, 1968). It has been also hypothesized that in these steps the coding capability of the nucleotides evolved from 2 to 4, that is, from being able to recognise only the pyrimidine-purine character of a nucleotide to the present recognition of the 4 bases (U,C,A,G).

The main problem of this evolutionary approach is that a change in the number of bases per codons breaks the normal reading frame. An accepted early possible solution to this problem assumes that codons have always had 3 bases but only the number of informative nucleotides inside a codon have passed from 1 to 2 and 3. The number 3 should be related to the stability of the codon anti-codon interaction. Crick himself proposed the first model of this kind (Crick 1968). Successively it has been proposed that di-nucleotides play a role in defining the evolution of the code. More recently, the hypothesis included the idea that two independent groups of di-nucleotides were at the origin of the genetic code: the so called prefix and suffix doublets (Wu et al., 2005). Another recent work that emphasizes the role of dinucleotides put the attention on primeval symmetries of dinucleotides, in particular, the possibility that ancient tRNAs were reversible and, thus, able to read a dinucleotide in both directions (Wilheim, 2004). From the biological point of view all these ideas are compatible with our mathematical model. In fact, another possible solution for the evolution of protein coding involves counting according the powers of 2. This would produce a smooth and plausible transition through duplication. As we will show, a possible scenario can lead to four base codons through duplication and to three base codons through information reduction. The idea that the genetic code evolved by reducing codon's dimension instead of augmenting it has been proposed recently (Baranov et al., 2009). On the other hand, our mathematical model suggests codons formed by pairs of dinucleotides. Thus, we can relax the length constraint and look for a first-principles explanation of the observed degeneracy. In fact, differently from (Wu et al., 2005) where prefix and suffix codons merge as to form a codon of length 3, we can couple two independent dinucleotides and form a length-4 codon. The main problem with this approach is that there are $4^4 = 256$ possible length-four codons whereas we know that the genetic code has 64 codons. Once again the issue can be solved by looking at the basic transformations and symmetries derived from first principles and described above. In practice, a doublet is coupled with another doublet obtained from one of the 4 transformations: Identity (I), Complement or Strong/Weak (C), Pyrimidine/Purine (Y/R), Rumer or Keto/Amino (R). In such a way we obtain 64 length-4 codons which we call *tesserae* (from the Greek τεσσερα = four). This points to primeval symmetries that may explain coding characteristics and degeneracy distribution of the genetic code. If we elaborate along this line of thought we can note that there are 4 basic symmetries associated to a DNA molecule:

- the identity (normal reading of a triplet in the coding strand in the 3'-5' direction).
- reverse complementary (normal reading of a triplet in the complementary strand in the 3'-5' direction).

- reverse (reading a codon in the coding strand in the reverse 5'-3' direction).
- complementary (reading a codon in the complementary strand in the reverse 5'-3' direction).

These symmetries themselves allow for coding amino acids with some degeneracy. Suppose, that a given tRNA is able to read a codon in both directions of the coding strand (see Fig. 1). In such a case, palindromic codons, (i.e., codons of the form XNX) can be associated to a single amino acid, but asymmetric codons of the form XNY, form pairs of two codons that code the same amino acid, i.e., XNY, and YNX (see Fig. 2). In this way, a form of degeneracy related to the primeval symmetries of codons arises naturally. Note that, counterintuitively, more symmetric codons correspond to less degenerate amino acids.

We can guess that by using the symmetries it should be possible to explain the degeneracy of the code with length-3 codons. However two problems arise: first the reverse complementary symmetry cannot produce the degeneracy; in fact, a codon formed by 3 different nucleotides cannot coincide with its reversed complement. In general, it can be proven that codons with an odd number of nucleotides cannot have the reverse complementary symmetry. On the contrary, codons with an even number of nucleotides can have it. The second problem is that the number of direct-reverse symmetric codons do not suffice for explaining the number of amino acids with lesser degeneracy. However, a complete description of the degeneracy from first principles can be achieved by using our tessera code. Now we show that we have 16 tesserae that possess the reverse symmetry and 16 that possess the reverse complementary property. In fact, palindromic words can be generated by coupling the 16 doublet with their reverse, see the first column of Table 8. In order to form the reverse complementary ones, we do the same with the reverse complementary transformation; this part of the tessera code is shown in the second column of Table 8. Now a fundamental observation can be done: one half of the mitochondrial genetic code can be obtained by using symmetric reversible adaptors (tRNAs) that act on the tesserae subset shown in Table 8 (32 over 256). It is possible to achieve this by using tRNAs that possess either the reverse or the reverse complementary symmetry. The solution that uses the reverse symmetry is simpler from the biological point of view and has been actually proposed in Wilheim (2004). The second solution implies the presence of an independent anticodon in the tRNA molecule whose sequence coincides with the codon (tessera), i.e. the anti-(anti-codon). Another possibility allows the implementation of the reverse-complementary symmetry without further assumptions on the tRNA primeval molecules: the reading of the complementary fibre of DNA. However this hypothesis seems to be in contradiction with an RNA origin of life.

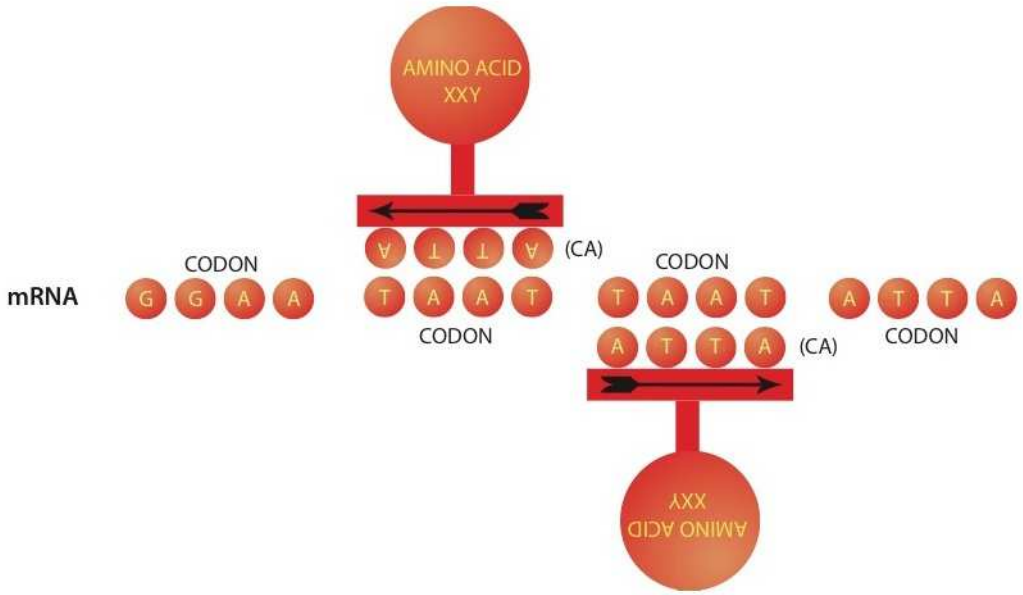


Figure 1: *A symmetric tRNA adaptor that can read mRNA in both directions. When the adaptor reads palindromic codons it recognizes the same codon (tessera), in the example TAAT.*

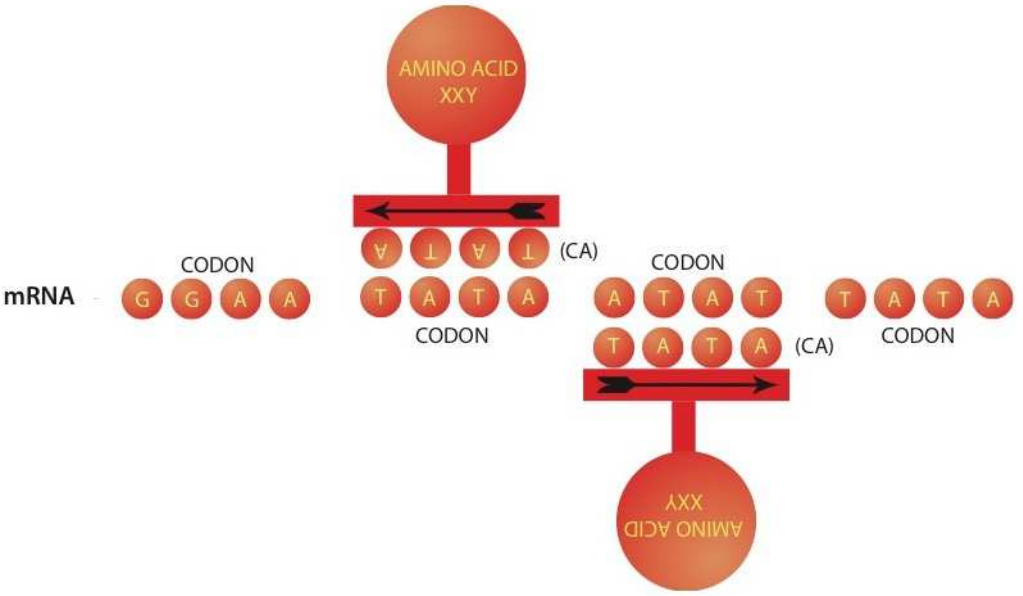


Figure 2: *A symmetric tRNA adaptor that reads self-complementary codons; it can recognize two of them, in the example TATA, and ATAT.*

	A	A	A	A	I	A	A	U	U	C
	U	U	U	U		U	U	A	A	
	C	C	C	C		C	C	G	G	
	G	G	G	G		G	G	C	C	
	A	U	U	A	C	A	U	A	U	I
	U	A	A	U		U	A	U	A	
	C	G	G	C		C	G	C	G	
	G	C	C	G		G	C	G	C	
	A	G	G	A	Y/ R	A	G	C	U	R
	U	C	C	U		U	C	G	A	
	C	U	U	C		C	U	A	G	
	G	A	A	G		G	A	U	C	
	A	C	C	A	R	A	C	G	U	Y/ R
	U	G	G	U		U	G	C	A	
	C	A	A	C		C	A	U	G	
	G	U	U	G		G	U	A	C	

Table 8: *4-base tesserae exhibiting the palindromic symmetry, (left), and the self-complementary symmetry, (right). In white columns we indicate the type of global transformation between the first 2 bases (first doublet) and the second ones (second doublet).*

In Table 9 we show the reading of the 32 symmetric codons listed in Table 8 with 24 symmetric adaptors that can read in both directions. It is clear that such adaptors can read only one palindromic codon, for example, the adaptor TTTT can read only the codon AAAA. However, reverse-complementary codons can be read in pairs by a reversible adaptor, for example, the adaptor AATT can read both TTAA and AATT (see also Fig. 2). The 8 tesserae on the upper left as well as the 8 at the bottom right are non degenerate, i.e., they can be read by only one adaptor (degeneracy 1). The central part of the table, instead, is formed by 8 pairs of reverse-complement codons. Any pair of reverse-complement codons can be read by the same reversible adaptor (degeneracy 2).

	A	A	A	A	I					
	C	C	C	C						
	A	U	U	A	C					
	C	G	G	C						
	A	G	G	A	Y/					
	C	U	U	C	R					
	A	C	C	A	R					
	C	A	A	C						
	A	A	U	U	C	U	U	A	A	C
	C	C	G	G		G	G	C	C	
	A	U	A	U	I	U	A	U	A	I
	C	G	C	G		G	C	G	C	
	A	G	C	U	R	U	C	G	A	R
	C	U	A	G		G	A	U	C	
	A	C	G	U	Y/	U	G	C	A	Y/
	C	A	U	G	R	G	U	A	C	R
						U	U	U	U	I
						G	G	G	G	
						U	A	A	U	C
						G	C	C	G	
						U	C	C	U	Y/
						G	A	A	G	R
						U	G	G	U	R
						G	U	U	G	

Table 9: *The 32 tesserae of Table 8 can be ordered according to their degeneracy as if they were read by 24 reversible adaptors. This describes half the degeneracy of the vertebral mitochondrial genetic code from first principles alone, compare with Table 7. The same can be obtained by using adaptors with two anti-codons.*

Observe that the distribution of degeneracy in this case is:

Degeneracy	# of tesserae
1	16
2	8

Table 10: *Degeneracy distribution for one half of the primeval mitochondrial genetic code obtained by reading all the symmetric tesserae (palindromic and reverse-complementary) by means of reversible tRNA adaptors.*

Hence, we have a biological description for half of the degeneracy of the mitochondrial genetic code, so that if we multiply by 2 the degeneracy numbers in Table 10 we obtain the full degeneracy of the mitochondrial genetic code.

How can we generate the other half of the code? Once again, the key is to study the group of transformations that links the two doublets. In fact, as shown in Gonzalez et al. (2008), all the possible global transformations of the nucleotides form a Klein V discrete group. These transformations are isomorphic to the symmetries of a rectangle:

In the first column of Table 8 we have listed all the palindromic tesserae; in the second one, all the reverse complementary ones. Also, between the first two letters (di-nucleotide) of a tessera and the two last ones, all the transformations are included in the two columns of primeval symmetries but in a different order (see the columns on the right of the tesserae). This table can be completed by implementing the same four group of transformations in the two remaining possible arrangements. In this way we obtain the set depicted on Table 11. Now, an elegant biological solution for the explanation of the degeneracy distribution of the mitochondrial genetic code can be realized simply by merging the two possibilities of symmetric primeval adaptors, i.e., including reverse symmetry and reverse-complement one, at the same time (see Fig. 3). In fact, if we apply these adaptors to the set generated in this way we obtain the complete degeneracy of the mitochondrial genetic code.

Summarizing, we have 8 primeval adaptors that can read 16 codons in pairs (degeneracy 2 of palindromic codons), 8 adaptors that can read other 16 codons in pairs (degeneracy 2 of reverse-complementary symmetric codons), and 8 adaptors that can read 32 codons grouped in quartets, the last two columns of Table 11. In fact, a reverse-complementary and reversible tRNA that acts on a codon of this kind is able to read four different codons associated to the symmetry transformations (see Figure 4). Observe that a valid tessera together with its complement always code for the same amino acid. Moreover, as in most variants of the code, we have two groups of degeneracy 2 corresponding to the stop signals, thus, we need only 24-2=22 adaptors for implementing the mitochondrial genetic code. Remarkably, this is exactly the number of tRNA adaptors in the vertebral mitochondrial code. Such result highlights the strong connection between non-power representations, group theory, and the possible biological implementation.
Interestingly, the degeneracy of the genetic code can be put in physical terms, that is, it can be associated to the adaptor's codon-anticodon bound energy. In fact, a completely symmetric adaptor under axial 180° rotations will exhibit the same bounding energy in the 3'-5' direction than in the 5'-3' direction. The bounding energy will be the same both for a codon and its reverse. If the adaptor possesses also a perfect symmetry under the orthogonal 180° rotation that exchange an anticodon with its complementary version, the bounding energy with a codon or with its reversed complement will be the same. Applying again an axial 180° rotation we obtain that also such bounding energy will be identical for a codon and its complementary (reversing the complementary reversed codon). Thus, the degeneracy of the code can be explained as a bounding energy degeneration between the different codons that code an amino acid and the unique primeval symmetric adaptor that carries such amino acid. Degeneracy here conserves its original meaning related to the existence of different quantum states that share the same energy. In extant translation systems the above symmetries should be detected as a blueprint of primeval translation systems.

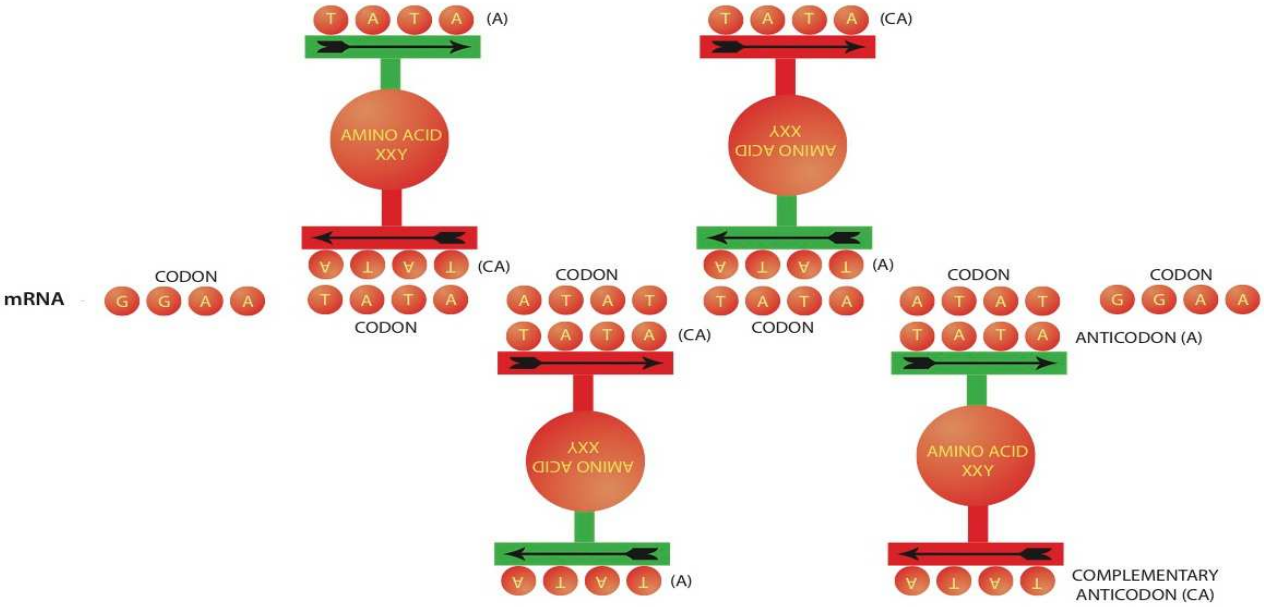


Figure 3: *reverse-complementary and reversible tRNA adaptors that read a codon and its complement on palindromic or self-complementary tesserae (degeneracy 2 amino acids).*

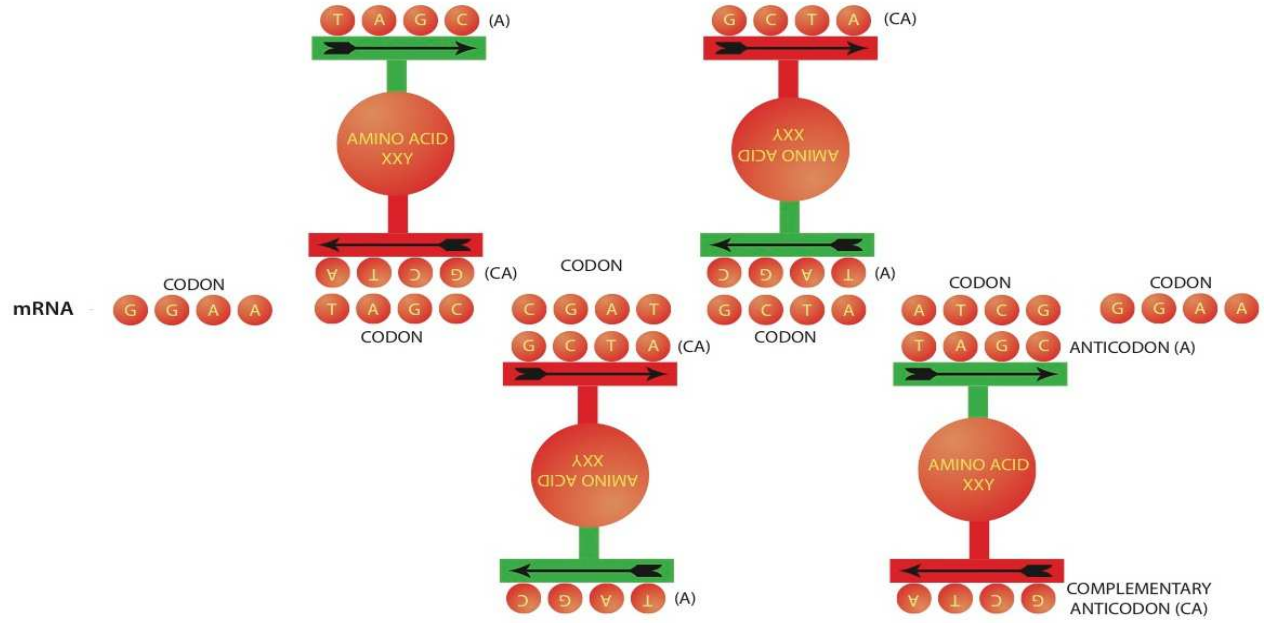


Figure 4: *reverse-complementary and reversible tRNA adaptors that read asymmetric tesserae (two last columns of Table 11). The tRNA can read four different tesserae so that the corresponding amino acid has degeneracy four.*

In fact, many proteins are coded in an anti-sense direction (very usual in mitochondria). Moreover, recent findings show the existence of a Chargaff rule (Chargaff, 1952; Chargaff et al., 1952) at the level of codons, supporting the hypothesis that in origin a codon and its complementary reversed version coded for the same amino acid (Jean Claude Perez, 2010).

	A	A	A	A	I	A	A	U	U	C	A	A	G	G	Y / R	A	A	C	C	R
	U	U	U	U		U	U	A	A		U	U	C	C		U	U	G	G	
	C	C	C	C		C	C	G	G		C	C	U	U		C	C	A	A	
	G	G	G	G		G	G	C	C		G	G	A	A		G	G	U	U	
	A	U	U	A	C	A	U	A	U	I	A	U	C	G	R	A	U	G	C	Y/ R
	U	A	A	U		U	A	U	A		U	A	G	C		U	A	C	G	
	C	G	G	C		C	G	C	G		C	G	A	U		C	G	U	A	
	G	C	C	G		G	C	G	C		G	C	U	A		G	C	A	U	
	A	G	G	A	Y / R	A	G	C	U	R	A	G	A	G	I	A	G	U	C	C
	U	C	C	U		U	C	G	A		U	C	U	C		U	C	A	G	
	C	U	U	C		C	U	A	G		C	U	C	U		C	U	G	A	
	G	A	A	G		G	A	U	C		G	A	G	A		G	A	C	U	
	A	C	C	A	R	A	C	G	U	Y / R	A	C	U	G	C	A	C	A	C	I
	U	G	G	U		U	G	C	A		U	G	A	C		U	G	U	G	
	C	A	A	C		C	A	U	G		C	A	G	U		C	A	C	A	
	G	U	U	G		G	U	A	C		G	U	C	A		G	U	G	U	

Table 10: *Complete 4-base (tessera) mitochondrial genetic code The white columns report the global transformation that generate a tessera by acting on the first dinucleotide..*

Table 10 highlights a quantity of striking properties of this code; the most important ones relate to its error-correction capability. For example, any arbitrary point mutation that affects a nucleotide in any position of any tessera leads to an invalid tessera. This property matters because it might be a fundamental key for the selection of such a code from an evolutionary point of view. Divergences of the code from Table 11 tend to be eliminated because variants are prone to incorporate errors by point mutations. Moreover, the coding of a protein can be made insensible to ±1 frame shifts. In fact, the unique way for having a valid tessera at step ±1 is to put the initial letter at the end. As the minimal degeneracy of an amino acid is 2 we have at least 2 different choices for appending a tessera, and thus, we can always append a tessera that starts with a different letter than the previous one making the chain immune to ± 1 frame-shift errors.

5) Conclusions

In this paper we have shown that the global degeneracy distribution of the vertebrate mitochondrial genetic code can be described by a mathematical model based on number representation systems. Remarkably, the mathematical theory that underpins the representation is the same that allows to describe the nuclear genetic code. Moreover, we have shown that the representation of the nuclear code can be obtained from a break in the symmetric structure of the mitochondrial code. In this respect, this article is a step forward towards a unified mathematical theory of the management of genetic information. The results also pave the way to a series of fundamental questions, which we do not pursue here, on the descendance of the nuclear code from the mitochondrial code.

The mathematical approach allows the uncovering of many different symmetries and anti-symmetries of the genetic code that seems strongly related to the organization of genetic information. Moreover, a guideline of the present work has been the search for possible error detection/correction mechanisms acting at the level of protein synthesis. The more symmetric form of the mitochondrial genetic code and the fact that the quantity of information contained in mitochondria is several orders of magnitude smaller than such contained in the nucleus, points to the presence of simpler mechanisms for error detection/correction in mitochondria. Thus, the study of the mitochondrial genetic code allows also to gain information on the nuclear strand. In particular, we are able to hypothesize the existence of a primeval mitochondrial genetic code composed by codons of four nucleotides that we have called *tesserae.* We found that such primeval genetic code used only 64 of the 256 available tesserae. Such 64 tesserae possess precise symmetry properties related to the discrete Klein V group of symmetries, and in particular, they are completely immune to arbitrary point mutations, that is, arbitrary changes in one letter of the tessera alone. Such property is strongly connected to usual error detection/correction mechanism, as such used in man-made technological systems for ensuring faithful transmission of digital data along noisy channels. Thus, the present work opens the door to the attack to the problem of error detection/correction in protein synthesis from a formal point of view strongly related to coding and communication theory.

Surprisingly, the four-letter codon hypothesis leads to a possible biochemical explanation for the origin of protein coding. In fact, the degeneracy of the vertebrate mitochondrial genetic code can be described exactly by assuming the existence of primeval tRNA adaptors with definite symmetry properties, such as reversibility and complementarity. In addition, the hypothesis explain naturally why such degeneracy can be described by using only 22 primeval tRNA adaptors. Recently, it has been proposed, that the genetic code might have originated from codon size reduction (Baranov et al., 2009). Moreover, an orthogonal ribosome (ribo-Q1) that decodes efficiently a set of 4-base codons has been built in a laboratory (Neumann et al., 2010). Hence, it is possible to show that all the suggestions and predictions that stem from the mathematical model have a plausible sound biological base.

Further investigations will include the study of the existence of a mapping that allow to pass from the tessera code to the present mitochondrial code. Ideally such function should preserve the good properties of the tessera code, including that of error detection and correction. A related topic is that of the role played within the tessera code by dichotomic classes, quantities that arise naturally from the mathematical representation of the genetic codes observed in nature.

Acknowledgements

We would like to thank Alberto Danielli and Julyan Cartwright for useful discussions and suggestions.

Bibliography

P.V. Baranov, M. Venin, G. Provan, Codon Size Reduction as the Origin of the Triplet Genetic Code, PLoS ONE 4(5): e5708. doi:10.1371, (2009).

ED, Chargaff, On the deoxyribonucleic acid content of sea urchin gametes, *Experientia* **8** (4): 143–145, (1952).

ED, Chargaff, Lipshitz R, Green C, Composition of the deoxypentose nucleic acids of four genera of sea-urchin". *J Biol Chem* **195** (1): 155–160, (1952).

FHC. Crick, The origin of the genetic code. *J Mol Biol.* **38**:367–379, (1968).

FHC Crick, Codon–anticodon pairing: The wobble hypothesis; *J. Mol. Biol.* **19** 548–555, (1966).

G. Gamow, Possible Relation between Deoxyribonucleic Acid and Protein Structures, Nature, vol.173, pp. 318 (1954).

Giannerini, S., Gonzalez, D.L., Rosa, R., 2012, DNA, circular codes and dichotomic classes: a quasicrystal framework, Philosophical Transactions of the Royal Society A, forthcoming.

Gonzalez, D.L., 2004. Can the genetic code be mathematically described? Med. Sci. Monit. 10 (4), 11–17.

Gonzalez, D.L., 2008. The mathematical structure of the genetic code. In: Barbieri M., Hoffmeyer J. (Eds.), The Codes of Life: The Rules of Macroevolution, vol.1. Springer, Netherlands, pp. 111–152 (Chapter 8).

Gonzalez, D.L., 2008. Error detection and correction codes. In: Barbieri, M., Hoffmeyer, J. (Eds.), The Codes of Life: The Rules of Macroevolution, vol. 1. Springer, Netherlands, pp. 379–394 (Chapter 17).

Gonzalez, D.L., Giannerini, S., Rosa, R., 2006. Detecting structure in parity binary sequences: error correction and detection in DNA. IEEE Eng. Med. Biol. Mag. 25, 69–81.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2008. Strong short-range correlations and dichotomic codon classes in coding DNA sequences. Phys. Rev. E 78 (5), 051918.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2009. The mathematical structure of the genetic code: a tool for inquiring on the origin of life. Statistica LXIX (3–4) 143–157.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2011, Circular codes revisited: A statistical approach,

Journal of Theoretical Biology 275, 21–28.

TH Jukes, Evolution of the Genetic Code from a Preceding Form, *Nature* 246, 22 - 26, (1973).

H. Neumann, K. Wang, L. Davis, M. Garcia-Alai, and J.W. Chin, Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome, Nature,Vol 464, doi:10.1038/08817 (2010).
R. D. Knight, L. F. Landweber, The early evolution of the genetic code. Cell, 101,no. 6, pp. 569 – 572, (2000).

J.C. Perez - Populations in Single-stranded Whole Human Genome DNA Are Fractal and Fine-tuned by the Golden Ratio 1.618, Interdiscip Sci Comput Life Sci 2: 1–13, (2010).

YB Rumer About the codon's systematization in the genetic code (in Russian), Proc Acad Sci U.S.S.R. (Doklady), 167: 1393, (1966).

T Wilhelm, and S. Nikolajewa, A new classification scheme of the genetic code, J. Mol. Evol., 59(5):598-605, (2004).

CR. Woese, *Proc. Natl. Acad. Sci. USA* 54:1546–52, (1965).

SA Wolfram, A New Kind of Science, Wolfram Media, Illinois, 2002

JT. Wong, A co-evolution theory of the genetic code. *Proc Nat Acad Sci USA.* **72**:1909–1912, (1975).

HL Wu, S Bagby, and JM van den Elsen, Evolution of the genetic triplet code via two types of doublet codons, J Mol Evol. 61(1):54-64, (2005).

E Zeckendorf, Representation des nombres naturels par un somme de nombres de Fibonacci ou de nombres de Lucas. Bulletin de la Societé Royale des Sciences de Liege, 41: 179-82, (1972).