Genes, Genomes and Genomics ©2007 Global Science Books

Degeneracy in the Genetic Code: How and Why?

Jean-Luc Jestin[1*] • Achim Kempf[2]

[1] Unité de Chimie Organique, URA CNRS 2128, Département de Biologie Structurale et Chimie, Institut Pasteur, Paris, France
[2] Department of Applied Mathematics, University of Waterloo, Ontario, Canada

Corresponding author: * **jjestin@pasteur.fr**

ABSTRACT

In the genetic code, which is nearly universal among all known organisms, most amino acids are coded for by more than one codon. For example for half of the genetic code's sixty-four codons, the corresponding amino acid is independent of the codon's third base. Interestingly, this degeneracy of the genetic code clearly reduces the deleterious effects of base substitutions at the third codon base. The genetic code possesses a significant number of further degeneracies and the question arises what the structure of the genetic code may be able to tell us about the circumstances of the genetic code's evolutionary origin. Here, we review selected articles in this context, beginning with works on the likely relatively recent origin of the small differences in the present-day genetic codes. We then review work that uncovered characteristic patterns in the genetic code, followed by work that describes the underlying symmetries in terms of mathematical models. Finally, we address the question what the structure of the genetic code might be able to tell us about the circumstances of the genetic code's evolutionary origin and therefore about the origin of life itself.

Keywords: codon, triplet, symmetry, coding, evolution

CONTENTS

INTRODUCTION

The genetic code establishes the correspondence between one or several codons and amino acids (Jones 1966). While the genetic code is quasi-universal for all known organisms, its pattern of degeneracies is, interestingly, more universal than the genetic code itself. For example, the mitochondrial genetic codes of vertebrates and of mollusks differ in the assignment of the AGR codons: these encode a stop signal in the former and the amino acid serine in the latter (**Fig. 1**) (Osawa 1992). Nevertheless, the degeneracy patterns of these two genetic codes are strictly identical. The genetic codes' pattern of degeneracies is however not fully universal. Amino acids are for example encoded by two, four or six codons in the mitochondrial code of vertebrates while amino acids are encoded by one, two, three, four or six codons in the standard genetic code.

In this article, we will summarize first evolutionary models that account for the diversity among the degeneracy patterns of genetic codes in section 1. Next, we will review characterizations of the degeneracy patterns by their symmetries in section 2. Subsequently, we will sum up mathematical models which efficiently describe the symmetry structure of the degeneracy of the codes in section 3. Finally, we will ask what may be learned from the structure of the genetic codes about the circumstances of the genetic code's evolutionary origin.

Evolutionary models accounting for the diversity of genetic codes

We begin by noting that progress has been made regarding the understanding of the diversification of the genetic code, i.e., regarding the likely causes of the small perturbations that the code experienced after it had essentially frozen its evolution about 3.8 billion years ago (Osawa 1992, 1995; Ardell 2001). The models for the evolution of the genetic code which we summarize below each provide rationales for how certain specific codons changed their assignments and therefore also changed the code's degeneracy.

In the first model, a codon is assumed to have been very rarely used prior to its reassignment to another amino acid (Caron 1990; Sengupta 2005). A stop codon may for example loose its function in chain termination prior to its reassignment to an amino acid (Osawa 1988). This mechanism has been tested experimentally. A synthetic amino acid such as *O*-methyl-L-tyrosine was introduced in response to an amber stop codon by expression of a tRNA and an aminoacyl-tRNA-synthetase in *Escherichia coli* (Wang 2001). This approach provides a general mechanism for enlarging the genetic code at stop codons as well as at codons encoding amino acids (Santoro 2002; Kwon 2003).

In the second model, a codon is assumed to have had ambiguous functions before its reassignment to another amino acid or stop. A codon may for example encode two amino acids (Ninio 1986, 1990). Experimental evidence in favor of this model was recently obtained. A codon encoding valine was assigned to the two amino acids valine and cysteine in *Escherichia coli* by exerting an appropriate

Received: 27 February, 2007. Accepted: 16 April, 2007.

Invited Opinion Paper

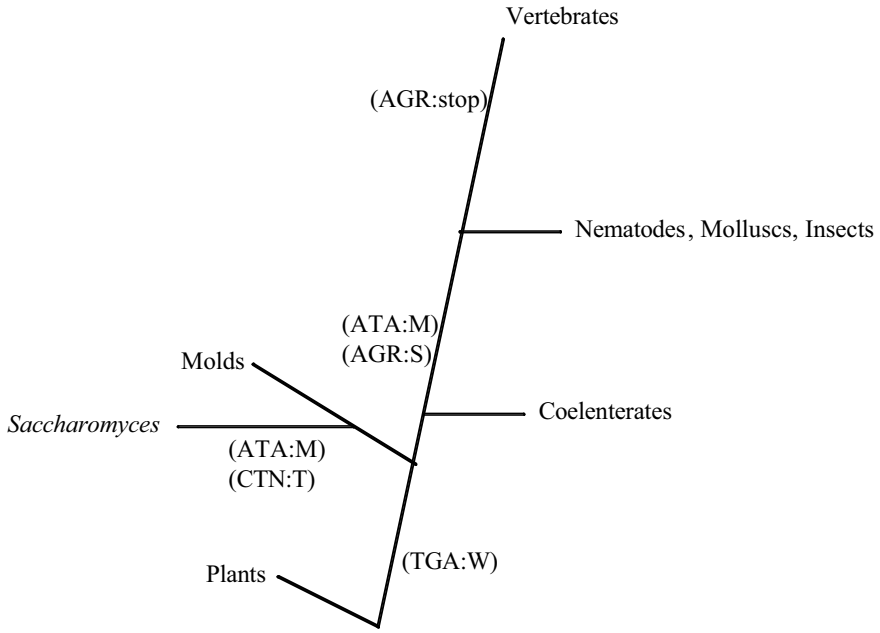

Fig. 1 Evolution of mitochondrial genetic codes from the standard universal code in fungi and animals (adapted from Jukes 1990).

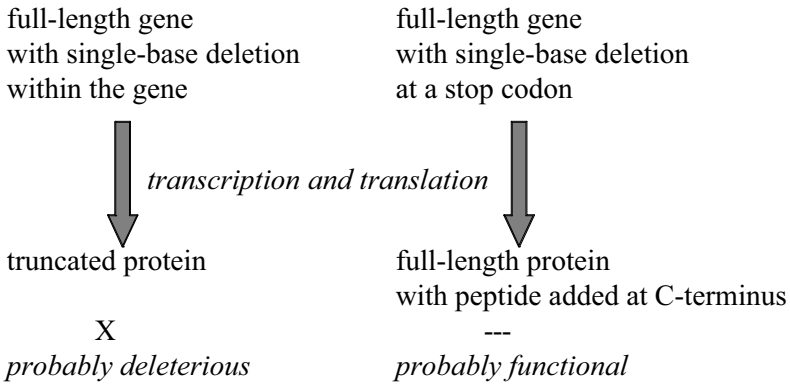

Fig. 2 Single-base deletions are less deleterious at stop codons than at codons encoding amino acids. (adapted from Jestin 1997).

selection pressure. In the absence of thymidine as a nutrient, the requirement of an active thymidilate synthase for cell growth allowed a valine codon to be read as a cysteine at a position which was essential for enzymatic activity (Döring 2001). Cys-tRNA^{Val} synthesis was catalyzed by valyl-tRNA-synthetase misaminoacylation: the synthetase was found to be mutated at its editing site (Döring 2001).

In the third model, the basic observation is that the assignment of codons to stop signals can be shown to be tightly linked to DNA polymerase fidelity (Jestin 1997). Sequences prone to single-base deletions are indeed less deleterious at the end of genes, i.e., at stop codons, than at codons encoding amino acids within genes (**Fig. 2**). Assigning sequences prone to single-base deletions to stop codons minimizes therefore the deleterious effects of these frameshift mutations. It was indeed reported that most single-base deletions induced by DNA polymerases occur at 5′-YTRV-3′ template sites opposite the purine R; mapping of the YTRV sequence on the codon table highlights four codons including the two stop codons TAR (Jestin 1997). This model is consistent with the theory stating that the genetic code has been evolutionarily optimized for error tolerance, i.e, to minimize the deleterious effects of mutations (Sonneborn 1965; Jestin 1997; Freeland 2000; Luo 2002; Dudkiewicz 2005). The close connection between the codon assignment of stop signals and sequence-dependent frameshifts catalyzed by DNA polymerases has an implication: different cellular compartments associated to polymerases with different fidelities are thus linked to genetic codes that may differ by their codon assignment of stop signals to minimize the deleterious effects of single-base deletions. There is evidence, therefore, that evolutionary models can account for modifications of the degeneracy pattern at a few codons between different genetic codes.

The degeneracy pattern of the genetic code characterized by its symmetries

We note that all the models that we mentioned so far concern the degeneracy in the genetic code independently of the physical and chemical properties of the four bases and the amino acids that are assigned to the various codons. Instead, the code's degeneracy is viewed solely as an intrinsic property of how the codons are grouped into sets of synonymous codons. Correspondingly, the observed symmetries of the code's degeneracy are viewed as intrinsic properties of the genetic code.

In this context, the very structure of the degeneracy pat-

Table 1 A representation of the genetic code.

Group I	Group II		
ACN T	CAY H	CAR Q	
TCN S	GAY D	GAR E	
CGN R	ATH I	ATG M	
GGN G	TTY F	TTR L	
CCN P	AAY N	AAR K	
GCN A	TAY Y	TAR -	
CTN L	AGY S	AGR R	
GTN V	TGY C	TGA -	TGG W

Representation of the standard genetic code in two groups depending on whether the third base of codons is necessary (group II) or not (group I) to define unambiguously an amino acid or a stop signal. Substitutions of G and C as well as T and A for the first base of codons leaves codons within the same group. Substitution of G and T as well as A and C for the three bases of a codon leads to a codon in the other group. Amino acids are indicated according to the single-letter nomenclature. A hyphen indicates a stop signal. Letters within triplets of bases follow the rules : N=A,T, C or G ; H=A,C or T ; Y=C or T ; R=A or G.

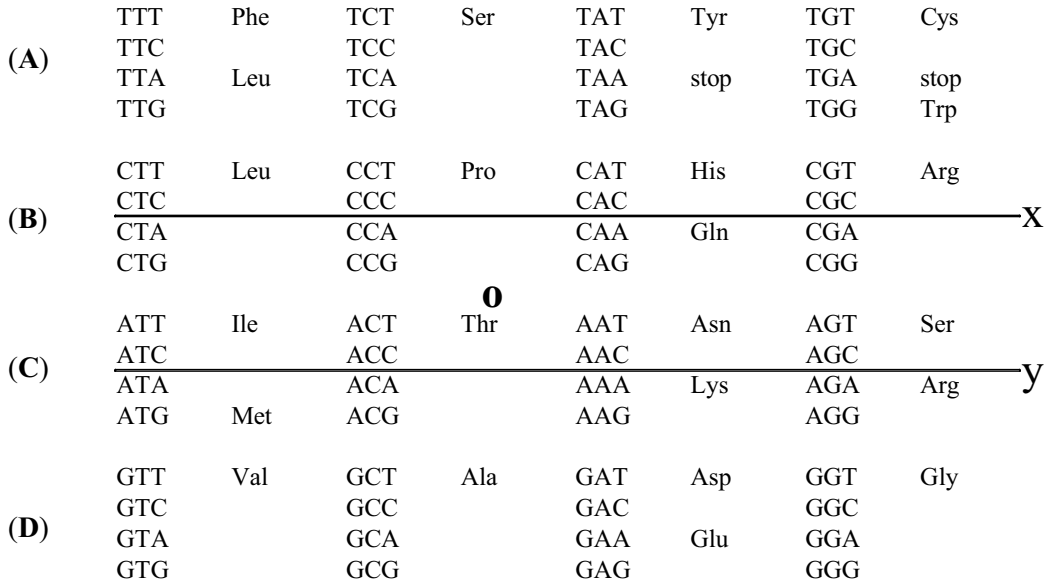

Fig. 3 Representation of the standard genetic code and its symmetries of degeneracy by base substitutions (adapted from Jestin 2006). The symbol **o** represents a central symmetry exchanging groups of four codons encoding a single amino acid into groups of four codons where the last base has to be defined to specify uniquely an amino acid acid or a stop codon. The four codon groups of lane (**A**) are exchanged into four groups of lane (**C**) by substitution of G and C as well as A and T at the first base of codons (symmetry of axis x). Similarly, lane (**B**) is exchanged into lane (**D**) by the same substitutions (symmetry of axis y). Groups of four codons encoding a single amino acid are then exchanged into groups of four codons with this same property. Similarly, groups of four codons requiring the last base to be defined to specify uniquely an amino acid or a stop codon, are exchanged into groups of four codons with this same property.

tern of the genetic code has been analyzed in detail, namely by studying its symmetries with respect to all possible base substitutions. This work has shown, in particular, that the genetic code can be split into two groups of 32 codons, one group in which knowledge of the third base of a triplet is necessary to determine the corresponding amino acid (or a stop signal) and one group in which the third base does not carry coding information (Rumer 1966; Danckwerts 1975). Rumer noted, in particular, the presence of a symmetry which exchanges both groups into each other (**Table 1**; **Fig. 3**). The symmetry exchanges T and G as well as A and C and is applied to all three bases of codons (Rumer 1966; Scherbak 1989). More recently, an additional symmetry which exchanges each group into itself was identified (Jestin 2006). This symmetry exchanges G and C as well as A and T at the first base of codons (**Fig. 3**). The corresponding mutations exchange codons for which the third base does not have to be specified (to define an amino acid unambiguously) into codons with this same property.

Mathematical models describing degeneracy in the genetic code

In the meantime, there is a body of work which explores the possibility that there may exist a simple mathematical structure that correctly summarizes all the symmetries in the degeneracy of the genetic code. Much of this work has been motivated by the example of elementary particle physics. There it was found over the past century that all known species of particles (such as electrons, muons, quarks etc.) can be grouped into so-called multiplets. This multiplet structure possesses a beautiful mathematical description in terms of representations of so-called compact Lie groups. The use of Lie groups is today at the heart of the standard model of particle physics as it has repeatedly proved its predictive power in experiments (Weinberg 1995). In this context, it is important to note that compact Lie groups are mathematical objects that are continuous. Nevertheless, the representations of compact Lie groups yield discrete structures, such as the discrete way in which particles are grouped into families and multiplets. The reason why compact Lie groups, while being continuous themselves, can yield discrete structures, is essentially the same as the reason why, for example, the vibrations of a continuous object such as a string produce a discrete frequency spectrum of overtones.

Here, in the case of the genetic code, the structure to be described mathematically is the genetic code's pattern of degeneracies, which is discrete. Motivated by elementary particle physics, the idea has been pursued that the genetic code's degeneracies may also be describable in terms of representations of certain Lie groups. The outcome has been that there exist three suitable Lie groups, though only with a caveat: the Lie groups must be "partially broken" (Hornos 1993; Forger 2000). The concept of broken Lie group symmetry is also used in elementary particle physics, though with less arbitrariness. Looking beyond regular Lie groups the authors also found that three particular so-called Lie super groups are able to reproduce the discrete degeneracy structure of the code.

In the search for a mathematical description of the genetic code's degeneracies, the net was cast even wider in (Frappat 2000, 2001). There, instead of the set of Lie groups, the set of so-called quantum groups was considered. The structure of their representations is discrete as well. The defining characteristic of quantum groups is that when representations of a quantum group are combined they behave in the same way that strands of hair do when being braided: there is a notion of over and of under crossing. Quantum groups carry several parameters, the main parameter usually being called q. In the limit $q \to 1$, the over and under crossing of the "strands" become indistinguishable and the quantum group reduces to a Lie group. The authors considered a certain quantum group in the limit $q \to 0$. This limit is mathematically singular but appears to yield a structure that closely matches that of the degeneracies of the genetic code, along with some features of the distribution of the corresponding amino acids' physical and chemical properties, such as hydrophobicity. Further, there has been work which has matched the structure of the genetic code, along with related physical and chemical properties of both the bases and the amino acids, to the structure of finite and therefore discrete groups. In this work, the so-called Klein group plays a prominent role (Antoneli 2004).

It is difficult to assess the relative significance of the various mathematical descriptions that have been found for the symmetry patterns of the degeneracies of the genetic code, such as the descriptions in terms of finite-, Lie- or quantum group symmetries. In elementary particle physics the emergence of Lie group symmetries has its well-understood origin in an overarching physical principle: particles possess so-called inner (or iso-spinor) degrees of freedom and the presence of the Lie group symmetry expresses the principle (the so-called gauge principle) that certain angles of those inner degrees of freedom are unobservable. Any particular choice of symmetry group leads to characteristic predictions and the correct symmetry Lie groups have been determined experimentally. In the case of the genetic code, the symmetries of its degeneracy pattern are clearly not random. Nevertheless, it is unclear if, analogously to elementary particle physics, there should exist some overarching biological principle which would have enforced that the genetic code evolved to possess a particular symmetry described by a Lie-, quantum or finite group. Instead, many of the symmetries in the genetic code may exist, for example, simply because the code started out as a smaller code which then became extended to triplets in a simple way. In order to establish that the match of the code's structure with a particular mathematical symmetry is not accidental, it would be necessary (though probably very difficult) to derive experimentally testable predictions that are specific to the symmetry group in question. It will be of interest to try and identify the minimal properties of the coding system that are sufficient to account for the symmetries in the genetic code.

Finally, also the question has been addressed what, if anything, may be deduced from the structure of the genetic code about the circumstances of the code's evolutionary origin (Gutfraind 2006). The starting point here has been the working assumption that very early in evolution, before the genetic code froze, the genetic machinery was still highly prone to errors and that, therefore, the genetic code's assignment of codons to amino acids was under strong evolutionary pressure towards the conservation of genetic information. In this scenario, the code evolved its pattern of degeneracies at least in part so as to make the most likely mutations neutral or of low impact to the phenotype. If so, as has been pointed out in (Gutfraind 2006), it should be possible on the basis of the genetic code's pattern of degeneracies to mathematically reconstruct which types of genetic errors were more and which were less frequent at that time. The first analysis of this type reported that the frequencies of genetic errors (and the frequencies of the nucleotides) were indicative of a likely non-thermophile origin of the code.

If indeed the genetic code evolved its pattern of degeneracies so that the phenotypic impact of the more frequent genetic errors was lessened by the code's degeneracies, then the presence of those well-studied symmetries in the degeneracies of the code indicates the presence of corresponding symmetries in the evolutionary pressures at the time that the code first formed. This opens up the possibility that mathematical findings about the symmetries of the genetic code, such as those that we reported above, can be translated into statements about the earliest evolutionary processes and the environment in which the genetic code first formed. We may not yet have fully decoded all the information that is contained in the genetic code.

REFERENCES

Antoneli F, Forger M, Hornos JEM (2004) The search for symmetries in the genetic code: finite groups. *Modern Physics Letters B* **18**, 971

Ardell DH, Sella G (2001) On the evolution of redundancy in genetic codes. *Journal of Molecular Evolution* **53**, 269-281

Caron F (1990) Eucaryotic codes. *Experientia* **46**, 1106-1117

Danckwerts HJ, Neubert D (1975) Symmetries of genetic code-doublets. *Journal of Molecular Evolution* **5**, 327-332

Döring V, Mootz HD, Nangle LA, Hendrickson TL, de Crécy-Lagard V, Schimmel P, Marlière P (2001) Enlarging the amino acid set of *Escherichia coli* by infiltration of the valine coding pathway. *Science* **292**, 501-504

Dudkiewicz A, Mackiewicz P, Nowicka A, Kowalczuk M, Mackiewicz D, Polak N, Smolarczyk K, Banaszak J, Dudek MR, Cebrat S (2005) Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. *Future Generation Computer Systems* **21**, 7, 1033-1039

Forger M, Sachse S (2000) Lie superalgebras and the multiplet structure of the genetic code. *Journal of Mathematical Physics* **41**, 5407-5422

Frappat L, Sorba P, Sciarrino A (2001) Quantum groups and the genetic code. *Theoretical and Mathematical Physics* **128**, 856-859

Frappat L, Sorba P, Sciarrino A (2000) A model of the genetic code *International Journal of Modern Physics B* **14**, 2485

Freeland SJ, Knight RD, Landweber LF, Hurst LD (2000) Early fixation of an optimal genetic code. *Molecular Biology and Evolution* **17**, 511-518

Gutfraind A, Kempf A (2007) Error-reducing structure of the genetic code indicates code origin in non-thermophile organisms. *Origin of Life and Evolution of Biospheres*, in press

Hornos JEM, Hornos YMM (1993) Algebraic model for the evolution of the genetic code. *Physical Review Letters* **71**, 4401

Jestin JL (2006) Degeneracy in the genetic code and its symmetries by base substitutions. *Comptes Rendus Biologies* **329**, 168-171

Jestin JL, Kempf A (1997) Chain-termination codons and polymerase-induced frameshift mutations. *FEBS Letters* **419**, 153-156

Jones OW, Nirenberg MW (1966) Degeneracy in the amino acid code. *Biochimica Biophysica Acta* **119**, 400-406

Jukes TH, Osawa S (1990) The genetic code in mitochondria and chloroplasts. *Experientia* **46**, 1117-1126

Kwon I, Kirshenbaum K, Tirrell DA (2003) Breaking the degeneracy of the genetic code. *Journal of the American Chemical Society* **125**, 7512-7513

Luo LF (2002) Construction of genetic code from evolutionary stability. *Biosystems* **65**, 83-97

Ninio J (1986) Divergence in the genetic code. *Biochemical Systematics and Ecology* **14**, 455-457

Ninio J (1990) The revised genetic code. *Origin of Life and Evolution of Biospheres* **20**, 167-171

Osawa S, Jukes TH (1988) Evolution of the genetic code as affected by anticodon content. *Trends in Genetics* **4**, 191-198

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. *Microbiology Reviews* **56**, 229-264

Osawa S (1995) *Evolution of the Genetic Code*, Oxford University Press, Oxford, 205 pp

Rumer YB (1966) About the codon's systematization in the genetic code. *Proceedings of the Academy of Sciences of the USSR* **167**, 1393-1394

Santoro SW, Wang L, Herberich B, King DS, Schultz PG (2002) An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nature Biotechnology* **20**, 1044-1048

Sengupta S, Higgs PG (2005) A unified model of codon reassignment in alternative genetic codes. *Genetics* **170**, 831-840

Shcherbak VI (1989) Rumer's rule and transformation in the context of the cooperative symmetry of the genetic code. *Journal of Theoretical Biology* **139**, 271-276

Sonneborn TM (1965) Degeneracy in the genetic code: extent, nature and genetic implications. In: Bryson V, Vogel HJ (Eds) *Evolving Genes and Proteins*, Academic Press, New York, pp 377-397

Wang L, Brock A, Herberich B, Schultz PG (2001) Expanding the genetic code of *E. coli*. *Science* **292**, 498-500

Weinberg S (1995) *The Quantum Theory of Fields I*, Cambridge University Press, Cambridge, 635 pp