Modern Physics Letters B, Vol. 18, No. 18 (2004) 971–978
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

THE SEARCH FOR SYMMETRIES IN THE GENETIC CODE: FINITE GROUPS*

FERNANDO ANTONELI JR.[†]
Department of Mathematics, University of Houston,
651 Philip Guthrie Hoffman Hall, Houston, TX 77204-3008, USA
antoneli@math.uh.edu

MICHAEL FORGER
Departamento de Matemática Aplicada, Instituto de Matemática e Estatística,
Universidade de São Paulo, Caixa Postal 66281, BR-05311-970 São Paulo SP, Brazil
forger@ime.usp.br

JOSÉ EDUARDO M. HORNOS
Departamento de Física e Ciências dos Materiais, Instituto de Física de São Carlos,
Universidade de São Paulo, Caixa Postal 369, BR–13560-970 São Carlos SP, Brazil
hornos@if.sc.usp.br

Received 1 April 2004

We give a full classification of the possible schemes for obtaining the distribution of multiplets observed in the standard genetic code by symmetry breaking in the context of finite groups, based on an extended notion of partial symmetry breaking that incorporates the intuitive idea of "freezing" first proposed by Francis Crick, which is given a precise mathematical meaning.

Keywords: Finite groups; representation theory; genetic code.

PACS No.: 87.10.+e

The purpose of this letter is to present results concerning the possibility of reproducing the multiplet structure of the standard genetic code, as shown in Table 1, through the procedure of symmetry breaking using finite groups. Combined with the analogous analysis using Lie algebras and Lie superalgebras carried out earlier, this completes the mathematical program of performing a full search for symmetries in the genetic code, as set forth in Ref. 1. Methodologically, the main outcome is the explicit formulation of a procedure of "partial" symmetry breaking which is more general than the traditional Goldstone–Higgs type mechanism, incorporating

*Work supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), Brazil.
[†]On leave from Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil.

Table 1. The standard genetic code.

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

the intuitive idea of "freezing" in the evolution of the genetic code first proposed in 1968 by Francis Crick.[2]

The genetic code is the table of codon to amino acid assignments that governs the process of protein synthesis in all living organisms. The genetic information for protein synthesis is stored in DNA and RNA in the form of sequences of four nucleic bases, namely adenine A, cytosine C, guanine G and thymine T (in DNA) or uracile U (in RNA), from where it is read in triplets called codons. Hence there are altogether 64 different codons which constitute the basic units of genetic information. On the other hand, proteins are synthesized from 20 different amino acids. The genetic code is the dictionary governing the process of information transfer, or translation, from the DNA/RNA language with 64 words to the protein language with 21 words (20 amino acids plus the "Stop" signal). It is degenerate in the sense that codons can be arranged into multiplets such that two codons belong to the same multiplet if and only if they code for the same amino acid, or the "Stop" signal. This gives rise to a decomposition of the 64 codons into three sextets (Arg, Leu, Ser), five quartets (Ala, Gly, Pro, Thr, Val), two triplets (Ile, "Stop"), nine doublets (Asn, Asp, Cys, Gln, Glu, His, Lys, Phe, Tyr) and two singlets (Met, Trp).

In the first decade after its definite establishment, the genetic code was generally believed to be universal — used by all living organisms. However, the discovery that mammalian mitochondria use a slightly different code, where the significance of the UGA codon is changed from "Stop" to Trp, and the subsequent identification of entire families of nonstandard codes, mitochondrial as well as nuclear, triggered renewed interest among biologists concerning the origin of the code and its properties. A phylogenetic analysis of these deviations revealed that these nonstandard codes are relatively recent (no more than about 1.5 billion years old) and have been formed from the standard code (whose origin dates back to about 3.8 billion years ago) by posterior deviations.[3,4] Understanding the formation of the nonstandard codes from the standard code and understanding the evolution of the standard code itself are therefore logically distinct and *a priori* independent problems, the latter being intimately tied up with the quest for understanding the origin of life on Earth.

The central idea of the algebraic approach proposed in Ref. 1, and exposed in detail in Ref. 5, is to view the distribution of multiplets found in the standard code as the result of a symmetry breaking process. Starting out from a 64-dimensional irreducible representation — or codon representation, for short — of some primordial symmetry group or algebra, the standard code has, according to this picture, evolved into its present form through a sequence of transitions, each of them accompanied by a reduction of the symmetry existing at the previous stage to an appropriate maximal subgroup or subalgebra. In the last step, this reduction is allowed to be partial, in the sense that some of the multiplets that would normally break up are allowed to remain intact, or "frozen."

This general strategy can be implemented in various algebraic categories and involves two distinct steps: (1) the determination of the codon representations, and (2) the analysis of their branching rules under reduction to subgroups/subalgebras or chains of subgroups/ subalgebras. Such a program was first carried out for Lie algebras[1] (see Refs. 5 and 6 for a detailed exposition) and later extended to Lie superalgebras.[7] Performing such a classification for finite groups, which is perhaps the most natural context for this kind of investigation, is however much more difficult and has for a long time remained a challenging open problem. In what follows, we shall present the central results of this analysis, which has been completed recently.[8]

Our starting hypothesis is that the primordial symmetry should be given by a simple finite group or by one of its satellites, which are its upward, downward or mixed extensions. The main tool used here is the classification theorem for simple finite groups, whose proof has been completed in the 1980's. (See Ref. 9 for a presentation of the pertinent theory.) Briefly, the simple finite groups fall into four different types: the cyclic groups $\mathbb{Z}_p$ of prime order p; the alternating groups Alt_n for $n \geq 5$; the 16 series of simple finite groups of Lie type, also known as the (untwisted or twisted) Chevalley groups; and finally the 26 sporadic groups, the largest of which is the famous monster. Their satellites include: (a) proper central extensions, or covering groups (upward extensions); (b) extensions by groups of outer automorphisms (downward extensions); and (c) mixtures of both. All three types of extension are well known from Lie group theory: for example, the first occurs when passing from $SO(3)$ to $SU(2)$ in order to include half-integer spin, whereas the second corresponds to extending $SO(3)$ to $O(3)$ in order to include parity. In particular, admitting central extensions allows the inclusion of projective representations.

Table 2. Number N_l of linear and N_p of projective codon representations of simple finite groups and their downward extensions: alternating and symmetric groups.

G	$\lvert G\rvert$	N_l	N_p
Alt_8	20.160	1	1
Alt_{10}	1.814.400	0	2
Alt_{14}	43.589.145.600	0	1
Alt_{15}	653.837.184.000	0	2^*
Alt_{65}	$65!/2$	1	0
$Sym_8 = Alt_8.\mathbb{Z}_2$	40.320	2	2
$Sym_{13} = Alt_{13}.\mathbb{Z}_2$	6.227.020.800	0	1
$Sym_{14} = Alt_{14}.\mathbb{Z}_2$	87.178.291.200	0	2^*
$Sym_{65} = Alt_{65}.\mathbb{Z}_2$	$65!$	2	0

The first task is the determination of all codon representations of the simple finite groups and their satellites. The main difficulty to overcome here is the establishment of sufficiently stringent cutoffs on the parameters for the infinite series of these groups, which requires the use of a combination of sophisticated theorems from finite group theory, some of which have only recently become available. The remaining cases can then be handled using the ATLAS,[10] which is the basic source of information on representations of simple finite groups and their satellites, as well as the computer program GAP,[11] which allows the calculation of character tables of arbitrary finite groups, up to a certain order. Details of this analysis will be published elsewhere,[12] here we just summarize the results. First, the cyclic groups can be discarded immediately because they are abelian, and inspection of the ATLAS shows that only one sporadic group appears, namely the second Janko group J_2: it has two inequivalent pseudo-real projective codon representations. For the alternating groups and their satellites, among which one finds the symmetric groups, the list of codon representations is given in Table 2; partial results in this direction can already be found in Ref. 14. Similarly, for the Chevalley groups and their satellites, the list of codon representations is given in Table 3. (In both cases, the symbol 2 indicates the presence of two independent codon representations, whereas the symbol 2^* stands for a pair of complex conjugate codon representations.)

The second task is the determination of the branching rules of all these codon representations, with a few exceptions, in order to see whether any of them, when reduced to an appropriate subgroup, will reproduce the multiplet structure of the genetic code. The exceptions concern the huge groups Alt_{65} and Sym_{65}, which have been excluded since, obviously, their codon representations can be broken so as to reproduce any distribution of multiplets whatsoever, as well as the large alternating and symmetric groups Sym_{13}, Alt_{14}, Sym_{14} and Sym_{15}, for which the computer calculations are unfeasible, mainly due to the lack of memory. For the remaining cases, the program has been fully implemented on a personal computer with the

Table 3. Number N_l of linear and N_p of projective codon representations of simple finite groups G and their downward extensions $G.A$: Chevalley groups

<table>
<tr><th>G</th><th>$|G|$</th><th>N_l</th><th>N_p</th></tr>
<tr><td>$A_2(4) = PSL_3(4)$</td><td>20.160</td><td>1</td><td>$1+2^*+2^*$</td></tr>
<tr><td>$B_2(3) = PSp_4(3)$</td><td>25.920</td><td>1</td><td>1</td></tr>
<tr><td>$^2B_2(8) = Sz(8)$</td><td>29.120</td><td>1</td><td>1</td></tr>
<tr><td>$^2A_2(4) = PSU_3(4)$</td><td>62.400</td><td colspan="2">1</td></tr>
<tr><td>$A_1(64) = PSL_2(64)$</td><td>262.080</td><td colspan="2">1</td></tr>
<tr><td>$A_1(127) = PSL_2(127)$</td><td>1.024.128</td><td>0</td><td>2^*</td></tr>
<tr><td>$B_3(2) = PSp_6(2)$</td><td>1.451.520</td><td>0</td><td>2^*</td></tr>
<tr><td>$G_2(3)$</td><td>4.245.696</td><td>2^*</td><td>0</td></tr>
<tr><td>$G_2(2) = {}^2A_2(3).\mathbb{Z}_2$</td><td>12.096</td><td colspan="2">1</td></tr>
<tr><td>$A_2(4).(\mathbb{Z}_2)_1$</td><td>40.320</td><td>2</td><td>2</td></tr>
<tr><td>$A_2(4).(\mathbb{Z}_2)_2$</td><td>40.320</td><td>2</td><td>$2+2^*+2^*$</td></tr>
<tr><td>$A_2(4).(\mathbb{Z}_2)_3$</td><td>40.320</td><td>2</td><td>$2+2^*+2^*$</td></tr>
<tr><td>$A_2(4).\mathbb{Z}_3$</td><td>60.480</td><td>$1+2^*$</td><td>0</td></tr>
<tr><td>$A_2(4).\mathbb{Z}_6$</td><td>120.960</td><td>$2+2^*+2^*$</td><td>0</td></tr>
<tr><td>$B_2(3).\mathbb{Z}_2$</td><td>51.840</td><td>2</td><td>2^*</td></tr>
<tr><td>$^2B_2(8).\mathbb{Z}_3$</td><td>87.360</td><td>$1+2^*$</td><td>0</td></tr>
<tr><td>$^2A_2(4).\mathbb{Z}_2$</td><td>124.800</td><td colspan="2">2</td></tr>
<tr><td>$^2A_2(4).\mathbb{Z}_4$</td><td>249.600</td><td colspan="2">$2+2^*$</td></tr>
<tr><td>$A_1(64).\mathbb{Z}_2$</td><td>524.160</td><td colspan="2">2</td></tr>
<tr><td>$A_1(64).\mathbb{Z}_3$</td><td>786.240</td><td colspan="2">$1+2^*$</td></tr>
<tr><td>$A_1(64).\mathbb{Z}_6$</td><td>1.572.480</td><td colspan="2">$2+2^*+2^*$</td></tr>
<tr><td>$G_2(3).\mathbb{Z}_2$</td><td>8.491.392</td><td>2^*+2^*</td><td>0</td></tr>
</table>

help of GAP. The first and most tedious step is the calculation of the lattice of subgroups for each of the pertinent groups, up to conjugacy. Due to the structure of the algorithm used by GAP in this calculation, it turns out that — in contrast to the situation prevailing for Lie algebras and Lie superalgebras — nothing is to be gained by restricting to maximal subgroups, so it is at this stage more efficient to disregard chains of maximal subgroups and instead proceed directly to the final subgroup or, when "freezing" is involved, to the pair of subgroups formed by the penultimate and the final subgroups in the chain, calculating the corresponding branching schemes. Details will be published elsewhere,[13] but the final result is surprisingly simple. As in the case of Lie algebras and Lie superalgebras, it turns out that there is no subgroup which would reproduce the distribution of multiplets found in the genetic code by symmetry breaking in the traditional sense. However, when "freezing" is allowed, there appear three cases in which the correct branching can be obtained, in any one of the available (linear or projective) codon representations, through branching to any one of several subgroup pairs (H, K), provided the breaking from H to K is performed using an adequate freezing prescription:

- $G = B_2(3).\mathbb{Z}_2$, $\tilde{G} = Sp_4(3).\mathbb{Z}_2$: six branching schemes with $H = Q_8 : (\mathbb{Z}_3^2 : \mathbb{Z}_2^2)$ and six branching schemes with $H = Q_8 : (\mathbb{Z}_3^2 : \mathbb{Z}_2)$, with various choices for K,
- $G = B_3(2)$, $\tilde{G} = Sp_6(2)$: one branching scheme, with $H = (\mathbb{Z}_2^2.\mathbb{Z}_2^4) : \mathbb{Z}_3^2$ and $K = Q_8 : (\mathbb{Z}_3^2 \times \mathbb{Z}_2)$,
- $G = G_2(3)$, $\tilde{G} = G_2(3)$: six branching schemes, all with $H = Q_8 : (\mathbb{Z}_3^2 : \mathbb{Z}_2)$ and with various choices for K.

Here, $\tilde{G}$ is the "minimal" covering group of G needed to rewrite all projective representations of G that occur as linear representations of $\tilde{G}$. In all cases, the subgroups H and K are solvable and are constructed from the group Q_8 of unit quaternions, the symmetric group S_3 or the dihedral group D_{12} and the cyclic groups $\mathbb{Z}_2$ and $\mathbb{Z}_3$ by iteratively taking direct products (denoted by $G_1 \times G_2$), semidirect products (denoted by $G_1 : G_2$), and nonsplit extensions (denoted by $G_1 . G_2$); in particular, $\mathbb{Z}_p^n$ is an abbreviation for $\mathbb{Z}_p \times \ldots \times \mathbb{Z}_p$ (n factors) while the convention in the last two cases is that G_1 denotes the normal subgroup and G_2 the quotient.

In order to derive complete branching trees, the subgroup pairs (H, K) must be completed to descending chains of subgroups, each maximal in the previous one, that interpolate between G and H. This can of course be done in many different ways, but the final symmetry breaking patterns are often identical or very similar. In Fig. 1, we present as an example the pattern generated by symmetry breaking along the shortest chain for the smallest primordial symmetry group and which exhibits the smallest amount of freezing in the last step; this is the chain

$$G \supset M \supset H \supset K \tag{1}$$

where G is $Sp_4(3).\mathbb{Z}_2$, of order 103.680, M is its maximal subgroup of order 2.592 (for its structure, see the ATLAS), H is $Q_8 : (\mathbb{Z}_3^2 : \mathbb{Z}_2^2)$, of order 288, and K is $Q_8 : D_{12}$, of order 96. (The pattern is the same for all four codon representations.)

The methodology employed in the analysis of finite groups suggests a technical definition of an extended form of symmetry breaking which is only partial, allowing for a certain amount of "accidental degeneracies" in the final distribution of multiplets. Such a partial symmetry breaking is described by a group G with a given representation and a *pair* (H, K) of subgroups of G such that K is a maximal subgroup of H. Considering the decomposition of the original representation of G into irreducible representations of H, and then the further splitting into irreducible representations of K, some of the irreducible H-multiplets that would normally split into several irreducible K-multiplets are allowed to remain intact, or "frozen." The restriction we propose to impose on this phenomenon of (partial) freezing is that whenever the same H-multiplet occurs with multiplicity > 1, all of its copies should behave in the same way: either they all split or else they all remain unbroken. In other words, the alternative of freezing applies not to single multiplets, but rather to isotypic components. This is the rule that has been used in our analysis, for Lie

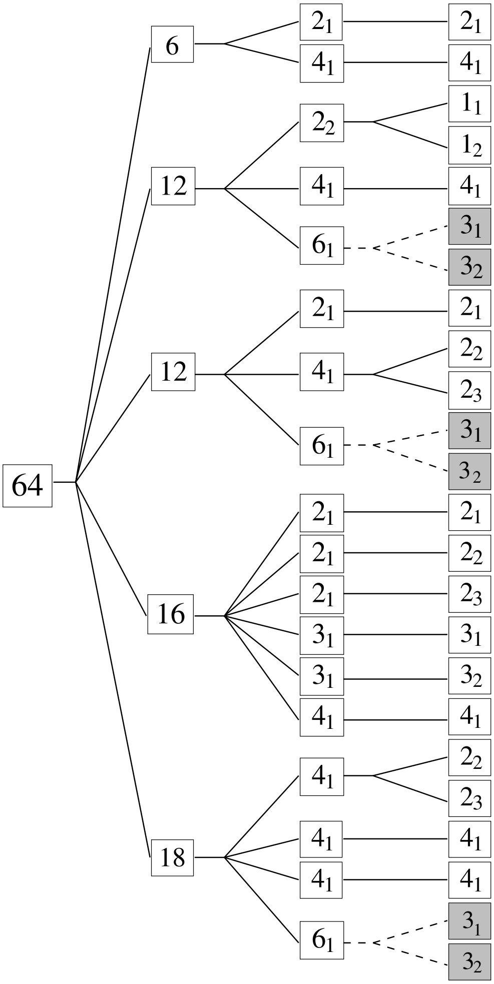

Fig. 1. Branching pattern for the codon representations of $G = Sp_4(3) . \mathbb{Z}_2$ along the chain (1): lower indices label irreducible representations of the same dimension with distinct characters.

algebras,[6] for Lie superalgebras,[7] and for finite groups.[8,13] We propose to call it the Higgs–Crick mechanism.

It would be interesting to find explicit examples of physical systems where symmetry breaking occurs with accidental degeneracies that fit into an extended notion of symmetry breaking such as the Higgs–Crick mechanism proposed here.

Summarizing, the question posed ten years ago[1] of whether there exists some symmetry principle underlying the degeneracy of the genetic code can now be completely answered. If one restricts the procedure of symmetry breaking to the traditional Goldstone–Higgs mechanism, there is no solution. However, with the extended Higgs–Crick mechanism formalized here, there are three Lie algebras, one Lie superalgebra, and three simple finite groups, that are able to generate the degeneracies of the code. Remarkably, symplectic algebras/groups appear in all three categories, suggesting strongly that the symplectic symmetry has been selected by evolution.

References

1. J. E. M. Hornos and Y. M. M. Hornos, Algebraic model for the evolution of the genetic code, *Phys. Rev. Lett.* **71** (1993) 4401–4404.
2. F. H. Crick, The origin of the genetic code, *J. Mol. Biol.* **38** (1968) 367–379.
3. S. Osawa, T. H. Jukes, K. Watanabe and A. Muto, Recent evidence for the evolution of the genetic code, *Microbiol. Rev.* **56** (1992) 229–264.
4. S. Osawa, *Evolution of the Genetic Code* (Oxford University Press, 1995).
5. J. E. M. Hornos, Y. M. M. Hornos and M. Forger, Symmetry and symmetry breaking: An algebraic approach to the genetic code, *Int. J. Mod. Phys.* **B13** (1999) 2795–2885.
6. F. Antoneli, L. Braggion, M. Forger and J. E. M. Hornos, Extending the search for symmetries in the genetic code, *Int. J. Mod. Phys.* **B17** (2003) 3135–3204.
7. M. Forger and S. Sachse, Lie superalgebras and the multiplet structure of the genetic code I: Codon representations, *J. Math. Phys.* **41** (2000) 5407–5422; Lie superalgebras and the multiplet structure of the genetic code II: Branching schemes, *J. Math. Phys.* **41** (2000) 5423–5444.
8. F. Antoneli, Grupos Finitos e Quebra de Simetria no Código Genético, Ph.D. Thesis, IME-USP, University of São Paulo (2003).
9. D. Gorenstein, *Finite Simple Groups* (Plenum Press, 1982).
10. J. H. Conway, R. T. Curtis, S. P. Norton, R. A. Parker and R. A. Wilson, *An Atlas of Finite Groups* (Clarendon Press, 1985).
11. The GAP Group, GAP — Groups, Algorithms, and Programming, Version 4.2; 1999 (http://www.gap-system.org).
12. F. Antoneli and M. Forger, Finite groups for the genetic code I: Codon representations, preprint RT-MAP 0006, IME-USP, University of São Paulo (2000).
13. F. Antoneli and M. Forger, Finite groups for the genetic code II: Branching schemes, in preparation.
14. R. D. Kent, M. Schlesinger and B. G. Wybourne, On algebraic approaches to the genetic code, *Canad. J. Phys.* **76** (1998) 445–452.