Algebraic Model for the Evolution of the Genetic Code

José Eduardo M. Hornos and Yvone M. M. Hornos

Departamento de Física e Ciência dos Materiais, Instituto de Física e Química de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-São Carlos, São Paulo, Brazil

(Received 22 June 1993)

A search for symmetries based on the compact simple Lie algebras is performed to verify to what extent the genetic code is a manifestation of some underlying symmetry. An exact continuous symmetry group cannot be found to reproduce the present genetic code. However, a unique approximate symmetry group is compatible with codon assignment for the fundamental amino acids and the termination codon. In order to obtain the actual genetic code, the symmetry must be slightly broken.

PACS numbers: 87.10.+e

The storage of genetic information in a cell is governed by deoxyribonucleic acid (DNA). The DNA molecules are large polymers composed of deoxyribonucleotides which contain a base, a sugar (deoxyribose), and a phosphate. The sugar and the phosphate groups are responsible for the well known helical backbone of DNA, and the bases sequence carries the genetic information. In DNA there are only four bases derived from purine and pyrimidine. The purines, adenine (A) and guanine (G), and the pyrimidines, thymine (T) and cytosine (C), form the double helix through the base pairs C-G and A-T bonded, respectively, by 3 and 2 hydrogen bonds. This pairing rule manifests itself not only by the spatial conformation on DNA, but also by the equal rate of cytosine to guanine and adenine to thymine. It is this pairing rule that makes one of the helices the exact template of the other one, so replication can be understood.

The transmission of information from DNA to protein building is a complex process of transcription and translation. In eucariotic cells DNA molecules are inside the nucleus of the cell and the proteins which they code are made outside of the nucleus in the citoplasma, more specifically in the ribosomes. The flow of information from DNA to the ribosomes requires another class of molecules, the ribonucleic acid (RNA) which are also constructed inside the nucleus. These molecules, rather than DNA, are the templates for protein synthesis; they leave the nucleus to the ribosomes to guide the synthesis.

RNA are unbranched polymers, much smaller than DNA, and are also composed of a sugar (ribose), a phosphate group, and a base. Different from DNA's, RNA's subdivide into classes, messenger *m*RNA, transfer *t*RNA, and ribosomal *r*RNA. While *t*RNA and *r*RNA are part of the protein-synthesizing machinery, *m*RNA's are the information carrying intermediates in protein synthesis. The size of RNA varies from as few as 75 to many thousands of nucleotides. *t*RNA's are smaller and they carry amino acids (a.a.) in an active form to the ribosome for peptide-bond formation in a sequence determined by the *m*RNA template. Ribosomal RNA's are the major component of ribosomes, but their precise role in protein synthesis is not yet known. The concept of *m*RNA was formulated by Jacob and Monod [1] in 1961. The genetic information from DNA is transcripted by RNA polymerases to *m*RNA. In a second step the information coded in the *m*RNA is translated into proteins. Here the idea of the genetic code emerges naturally. The code relates the sequence of bases in *m*RNA to the sequence of a.a. in a protein.

The beauty of the code lies in its simplicity. A sequence of three bases, called codon, specifies one a.a. The codons in *m*RNA are read sequentially by *t*RNA molecules in the ribosome. To form a codon we have four bases arranged in triplets, there are $4\times4\times4$ possible codons to code the 20 a.a., so the genetic code is degenerated. As soon as the code had been deciphered it was clear that a systematic investigation of the relation between the codons and the a.a. properties should be performed in order to understand the code.

A first attempt to explain the relation between the a.a. and the sequence of the bases was the proposal of a stereochemical theory by Woese [2]. This theory claims the existence of physicochemical interactions between the bases in the triplets and the a.a. which "fit stereochemically" with the particular combination. Another point of view was proposed by Crick [3] in the so-called frozen accident theory. In its extreme form the frozen accident theory implies that the allocation of codons to the a.a. at this point was entirely a matter of chance. In the same spirit Jukes [4] describes the evolution of the genetic code in terms of an optimization process in which the code evolved to a minimum, and the process was interrupted by a premature freeze in which the code was quenched in a local minimum. His analysis is based on the global efficiency of the protein production process rather than on quantitative arguments. The irregularities of the code are a testimony of the freeze. A different model was proposed by Wong [5,6] under the name of the coevolution theory. In this theory a few primordial a.a. existed, which have been specialized by biosynthetic processes forming new ones, finally reaching the contemporary code.

A second class of questions arises when we search for the relationships between the physicochemical properties of the a.a. and the code itself. Correlations between polarity, hydrophobicity, etc., and the codon assignments have been the subject of extensive investigations by many authors, see, for example, Di Giulio [7], Zimmermann, Eliezer, and Simba [8], and Jungck [9].

The main goal of the present paper is the search for symmetries in the genetic code and this leads us to the Lie group theory and the Cartan classification theorem. The algebraic approach has been used in different fields of science as high energy physics, the interactive boson model in nuclear physics, and the vibron model in molecular physics. The basic idea is to assume a fundamental group G and a chain of its subgroups. A dynamical process emerges assuming that the symmetry is successively broken throughout the chain. The possible symmetries have been classified. The four classical families SU(n), O($2n$), Sp(n), and O($2n+1$) and the exceptional groups G_2, E_4, E_6, E_7, and E_8 exhausted the simple groups. Here we not only report this search for the "best" underling symmetry, but also show that theory of symmetries can provide a mathematical framework in which a dynamical evolutionary pattern for the code emerges naturally. Rules for branching of the primitive a.a. into new ones are especially harmonious using this method.

From the results of this search a new class of models for the evolution of the genetic code based on the notion of symmetries has emerged. Group theory provides a set of well-defined simple rules for the construction of only a few "symmetric codes" among an enormous number of possibilities allowed by combinatorial analysis. Most of them give "non-sense" genetic codes, but only the one based on the sympletic symmetry is appropriate for genetics.

Let us now turn our attention to the code for mRNA shown in Table I. Some regularities are easily noted and a simple counting shows that there are three sextuplets, five quadruplets, two triplets, nine doublets, and two singlets.

Starting with the 64 codons and arranging different ways of distributing them among the 20 a.a. and one termination code, Bertman and Jungck [10] estimated that at least 10^{71} to 10^{84} different genetic codes like our contemporary one are possible. The central point in our analysis is that among this huge number of possible distributions of codons only a very limited number will correspond to Cartan symmetries and consequently generate an evolution pattern given by the group and its chains of subgroups. We look for groups which have a 64 dimension irreducible representation; this imposition leaves us with only SU(2), SU(3), SU(4), Sp(4), Sp(6), SO(13), SO(14), and G_2.

TABLE I. The genetic code.

First		Second position			Third
position	U	C	A	G	position
	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A
	Leu	Ser	Term	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

In order to obtain the present genetic code, we examine all chains of the 8 algebras and since the largest multiplicity in the code is 6, all higher dimension representations must be broken in representation of dimension smaller than or equal to 6.

A careful analysis of all the possible chains was made, including the SO(13) and SO(14), and it was found that there is no perfect symmetry conservation. Nevertheless the Sp(6) chain Sp(4)⊗SU(2) is the one that best reproduces the genetic code. This chain shows a "quenching" at the last step; i.e., the symmetry loss occurred at a point that can be detected very precisely in the last step of the evolution. It is interesting to emphasize here that the Sp(n) are highly noncommutative algebras, and the genetic code is also a noncommutative code, i.e., while ACU codes Phe, CAU codes His, and this is true for the entirety of the code. So the noncommutability of the code can be attributed to the fact that the starting group is Sp(6) and the chain leads via Sp(4)⊗SU(2) to the SU(2)⊗SU(2)⊗SU(2) symmetry.

The Sp(6) chain, with quenching term, that reproduces the degneracy of the genetic code and the perfect symmetry are displayed in Fig. 1, where each state is indicated by a horizontal bar with the multiplicity specified by the number over the bar; the horizontal axis indicates the sequence of the symmetry chain (evolution) as follows:

Step I: The Sp(6) [1,1,0] representation evolved to Sp(4)⊗SU(2), breaking the degeneracy of 64 into 6 different states with multiplicity 16, 4, 20, 10, 12, and 2; these states correspond to the primordial a.a. Step II: The Sp(4) representation breaking into SU(2)⊗SU(2). Step III: In the transition SU(2)⊃O(2) the second unitary group lifts the degeneracies as indicated in Fig. 1. Step IV: The last SU(2) breaks in O(2). It was at this stage of evolution that the freeze occurred. Step V: Figure 1(b) shows how the code would be if the symmetry breaking were allowed to continue. At this stage a perfect symmetry code could be written.

The analysis of Fig. 1 shows some aspects of the evolution of the genetic code. At step II the Sp(4)⊗SU(2) chain goes to SU(2)⊗SU(2)⊗SU(2), and the model resembles the product of three SU(2) groups, one for each base of the codon. This property is interesting if we are analyzing mutations. Also at this step the 6 primordial a.a. subdivide themselves into 14 a.a. with only one

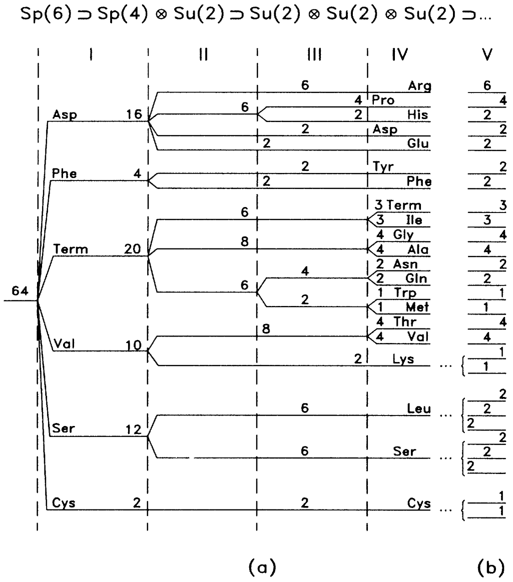

FIG. 1. (a) The genetic code multiplicity can be obtained from the Sp(6)⊃Sp(4)⊗SU(2)⊃SU(2)⊗SU(2)⊗SU(2)⊃SU(2)⊗Õ(2)⊗SU(2)⊃SU(2)⊗Õ(2)⊗O(2). The last SU(2) breaking is partial. A tentative assignment of the primordial a.a. is suggested in step I. (b) The total break of the last SU(2) in O(2). In this case Phe, Ser, Arg, and Cys would further subdivide leading to a code with 26 a.a. and the termination code.

of them, Cys, remaining unaltered. We note that Jukes [4], based on the study of *t*RNA's of various species, suggested an archetypal code containing 14 or 15 a.a. In the next step the symmetry is broken in the second SU(2), the Casimir operator used at this step is the square of L_z, and after this break a code with 16 a.a. results. The evolution of the code proceeds in IV, again breaking the last SU(2) in O(2). However, at this stage the symmetry undergoes a freeze. After Ala and Val are separated, Phe, Ser, Arg, and Cys would normally subdivide under the action of O(2), and instead they did not subdivide, freezing the code with only 20 a.a. "A symmetry perfect code" would have 27 a.a. Again Jukes suggested that the code should have generated 28 a.a. if the freeze had not occurred.

An operator that reproduces the multiplicities of the (standard) code is given by

$$\mathcal{H} = h_0 + h_1\mathcal{L}_4 + q_1L_1^2 + q_2L_2^2 + q_3L_3^2 + p_1L_{z_2}^2 + p_2(L_1^2+L_2^2)(D_3-3)L_{z_3}\,, \tag{1}$$

where h_0, h_1, q_1, q_2, q_3, p_1, and p_2 are arbitrary constants, $\mathcal{L}_4$ is the Sp(4) Casimir operator, L_1, L_2, and L_3 the angular momentum operators for each one of the three SU(2) groups, L_{z_2} and L_{z_3} the z components of the angular momentum operator of the O(2), and D_3 denotes the dimension of the last SU(2) representation. The eigenvalue of H when applied to a state assigned by the quantum numbers $|N_1,N_2,K_1,K_2,K_3,m_1,m_2,m_3\rangle$, where N_1,N_2 correspond to the representation of the Sp(4), K_1, K_2, and K_3 to the representation of the $SU_1(2)$, $SU_2(2)$, and $SU_3(2)$, respectively, and m_1, m_2, and m_3 to the representations of the three O(2), is given by

$$E = h_0 + h_1[(N_1+N_2)(N_1+N_2+4) + N_2(N_2+2)] + \sum_{i=1}^{3} q_i \tfrac{1}{4}(K_i)(K_i+2) + p_1m_2 + p_2\tfrac{1}{4}(K_1^2+K_2^2)(K_3-2)m_3\,, \tag{2}$$

where K_s^i are the eigenvalue of the A_1 algebra given by McKay and Patera [11]; the correspondence with the commonly used l_i is $l_i = K_i/2$.

At this point we have matched the multiplicities of the standard genetic code, but there is a large ambiguity in the assignment of the doublets, quadruplets, and sextuplets. The tentative assignment shown in step IV of Fig. 1 was done relating the eigenvalue given by Eq. (2) to the measured Grahtham polarities [12]; this was done by a minimization of the root mean square (rms) between calculated and measured polarities. The rms was calculated in the same spirit as that used for the IBM [13] and VIBRON [14] models. To the eigenvalues of the operator H we associated the polarities, and the correspondence of the group theoretical states to the a.a. is done after the minimization procedure considering the multiplicities.

In Table II we show the group theoretical assignment

TABLE II. Group theoretical assignment of the a.a. and polarities calculated by the present model.

	Group theoretical states	Polarities	
Name	N_1,N_2; K_1,K_2,K_3; m_1,m_2,m_3	Calc.	Expt.
Trp	$\|2,0; 0,2,1; 0,0,+\frac{1}{2}\rangle$	7.77	5.3
Met	$\|2,0; 0,2,1; 0,0,-\frac{1}{2}\rangle$	6.59	5.2
Cys	$\|0,0; 0,0,1; 0,0,\pm\frac{1}{2}\rangle$	3.92	4.8
Lys	$\|0,1; 0,0,1; 0,0,\pm\frac{1}{2}\rangle$	9.02	10.1
Asn	$\|2,0; 0,2,1; 0,\pm 1,+\frac{1}{2}\rangle$	8.77	10.0
Gln	$\|2,0; 0,2,1; 0,\pm 1,-\frac{1}{2}\rangle$	7.59	8.6
Tyr	$\|1,0; 0,1,0; 0,\pm\frac{1}{2},0\rangle$	5.68	5.4
Phe	$\|1,0; 1,0,0; \pm\frac{1}{2},0,0\rangle$	5.05	5.0
Asp	$\|1,1; 0,1,0; 0,\pm\frac{1}{2},0\rangle$	12.08	13.0
Glu	$\|1,1; 1,0,0; \pm\frac{1}{2},0,0\rangle$	11.45	12.5
His	$\|1,1; 1,2,0; \pm\frac{1}{2},0,0\rangle$	7.05	8.4
Ile	$\|2,0; 2,0,1; (\pm 1,0),0,+\frac{1}{2}\rangle$	5.60	4.9
Term	$\|2,0; 2,0,1; (+1,0),0,-\frac{1}{2}\rangle$	6.77	···
Val	$\|0,1; 1,1,1; \pm\frac{1}{2},\pm\frac{1}{2},-\frac{1}{2}\rangle$	5.30	5.6
Thr	$\|0,1; 1,1,1; \pm\frac{1}{2},\pm\frac{1}{2},+\frac{1}{2}\rangle$	5.89	6.6
Gly	$\|2,0; 1,1,1; \pm\frac{1}{2},\pm\frac{1}{2},-\frac{1}{2}\rangle$	8.45	7.9
Pro	$\|2,0; 1,1,1; \pm\frac{1}{2},\pm\frac{1}{2},+\frac{1}{2}\rangle$	7.86	6.6
Ala	$\|1,1; 1,2,0; \pm\frac{1}{2},\pm 1,0\rangle$	8.05	7.0
Leu	$\|1,0; 0,1,2; 0,\pm\frac{1}{2},(\pm 1,0)\rangle$	5.11	4.9
Ser	$\|1,0; 1,0,2; \pm\frac{1}{2},0,(\pm 1,0)\rangle$	5.74	7.5
Arg	$\|1,1; 2,1,0; (\pm 1,0),\pm\frac{1}{2},0\rangle$	6.68	9.1

Parameters

$h_0 = 3.88$; $h_1 = 0.64$; $q_1 = -2.70$; $q_2 = -2.2$
$q_3 = 0.03$; $p_1 = 1.00$; $p_2 = 1.18$; rms $= 1.314$

for the a.a., the values of the constants found in the minimization, and the calculated and experimental values of the polarities. This minimization was done in a very small phase space for the constants, but the agreement is good enough to show that this approach can be used to fit the other physicochemical properties of the a.a. It is of course a much more detailed study to verify which of the properties are best reproduced by our model; this will indicate the property that guided the evolution of the code in this symmetry path.

The assignment given in step I of Fig. 1 is based in the following arguments:

(1) If step IV corresponds to the present code and the termination signal was present in the primordial soup, then the 20-fold state in step I must be assigned to term.

(2) Since there are only two negatively charged a.a. at neutral *p*H (Asp on Glu), and both of them are in the same multiplet in step IV, we assigned Asp as a primordial a.a.

(3) From the three possible a.a. with multiplicity 6, Ser is easily formed in experiments designed to reproduce the original soup. Since the states with multiplicity 12 subdivide only in multiplets of dimension 6, Ser was assigned in the multiplet as primordial.

(4) The other three primordial a.a. have been assigned according to structural simplicity inside the multiplet.

Before our concluding remarks, we shall establish the limitations of the present approach. First it is not our intention to replace with our model a detailed microscopic biological, physical, and chemical analysis of the genetic code. Symmetry principles can and should be used only as a guide principle and a general framework in complement of a microscopic theory. Group theoretical principles will not tell which of the a.a. were primordial, but will only answer questions relating to the multiplicities and connectivities between different states (each state refers to one a.a.) with a defined multiplicity. Also the nonuniversality of the code can be analyzed in this formalism because different forms at breaking the symmetry will lead to different codes. Our program includes the study of other codes and we will search for the regularities among the models.

The use of symmetry principles to establish a spectrum generating algebra that reproduces the multiplicities of the genetic code can be viewed as a remarkable coincidence; and even a more drastic objection can be made remembering that there is no *a priori* justification for the use of symmetries in molecular biology or even in science. We stand on a technical basis: If there is any symmetry it should be found among Lie groups, discrete groups, or in the newly developed super and quantum groups. In this sense our search has the strength of a theorem, because all Cartan groups were studied. The other groups will be analyzed in a future work.

The main goal of the present paper is to introduce the group theoretical methods in the study of the genetic code. From this search a picture based in the Sp(4) $\times$SU(2) symmetry has emerged, and in general we establish the following: (1) From all the possibilities for the generation of the code and from the very small number of possible ways to arrive at a code guided by a dynamic breaking of a particular Cartan group, the genetic code can be obtained by immersion of the $\mathrm{Sp}(4)\otimes\mathrm{SU}(2)\supset\mathrm{SU}(2)\otimes\mathrm{SU}(2)\otimes\mathrm{SU}(2)$ in the Sp(6). (2) In the last step of the evolution the symmetry was "lost" in terms of its perfection but there is one way of breaking partially the symmetry so that the code is reproduced as known today. (3) A series of correlations concerning the evolution of a.a. with different multiplicities can be made.

The mathematical description of the a.a., in terms of a set of numbers, presented here and used to calculate the Grahtham [12] polarities, immediately suggests several topics for further investigations. For example, the full set of a.a. properties [Nakai, Kidera, and Kanchisa (1988), the database was kindly furnished to us by Peter Sibbald, EMBL] must be analyzed with this new model. A more ambitious project is to correlate the numbers that define each a.a. with important properties such as the folding.

Finally the newly found fact that there are only two different classes of syntheses [15] for the 20 a.a. must also be studied by this approach.

One of the authors (Y.M.M.H.) was supported by Conselho Nacional de Pesquisas (CNPq).

[1] F. Jacob and J. Monod, J. Mol. Biol. **3**, 318–356 (1961).

[2] C. R. Woese, D. H. Dugre, W. C. Saxinger, and S. A. Dugre, Proc. Natl. Acad. Sci. U.S.A. **55**, 966–974 (1966).

[3] F. H. C. Crick, J. Mol. Biol. **38**, 376–379 (1968).

[4] T. H. Jukes, J. Mol. Evol. **19**, 219–225 (1983); *Molecules and Evolution* (Columbia Univ. Press, Columbia, 1966).

[5] J. T. F. Wong, Proc. Natl. Acad. Sci. U.S.A. **73**, 2336–2340 (1976).

[6] J. T. F. Wong, Proc. Natl. Acad. Sci. U.S.A. **77**, 1083–1096 (1980).

[7] M. Di Giulio, J. Mol. Evol. **29**, 191–201 (1989).

[8] J. M. Zimmerman, N. Eliezer, and R. Simba, J. Theor. Biol. **21**, 170–201 (1968).

[9] J. R. Jungck, J. Mol. Evol. **11**, 211–224 (1982).

[10] M. O. Bertman and J. R. Jungck, Notices Am. Math. Soc. **25**, A-174 (1978).

[11] W. McKay and I. Patera, Table of Dimensions Indices and Branching Rules for Representations of Simple Lie Algebras (Marcel Dekker, New York, NY, 1981).

[12] R. Grahtham, Science **185**, 862 (1974).

[13] A. Arima and F. Iachello, *The Interacting Boson Model, Cambridge Monographs on Mathematical Physics* (Cambridge Univ. Press, Cambridge, 1987).

[14] J. E. Hornos and F. Iachello, J. Chem. Phys. **90**, 5284 (1989); F. Iachello and R. D. Levine, J. Chem. Phys. **77**, 3046 (1982).

[15] G. Eriani *et al.*, Nature (London) **347**, 203 (1990); R. Giege, J. D. Puglise, and C. Florentz (to be published).