Resenhas IME-USP 1996, Vol. 2, No. 4, 363 - 371.

Perspectives in the Algebraic Approach to the Genetic Code

J. E. Hornos and Y. M. M. Hornos

The main goal of this paper is to review the central features of the Algebraic Approach to the Genetic Code (AAGC) pointing the advances that are been worked today and the major open problems. The reader that is not familiar with the AAGC can find a brief description of the model in references [1] and [2], general comments in references [3] and [4] and a summary of main ideas of the model in reference [5]. A detailed review of the AAGC that contains a compact revision of the main biological and mathematical facts needs to understating the model is also in preparation [6].

The genetic code is the set of translation rules relating the sequence of chemical bases in RNA (ribonucleic acid) to the sequence of amino acids in proteins. The RNA molecules are large polymers composed by four bases, Adenine (**A**), Guanine (**G**), Uracil (**U**) and Cytosine (**C**) in a linear chain. Proteins are also chains formed by the twenty fundamental amino acids. The sequences of bases in RNA determine the sequence of amino acids in the protein. The major step in deciphering the genetic code was the discover by Crick et. al [7] that the genetic information was stored in triplets of bases, called codons, in RNA. Since we have four bases to be arranged in triplets there are 64 possible codons to code the twenty amino acids and a termination signal that indicates the end of protein formation. The genetic code can be viewed as a projective map from a set with 64 elements over a set that contains the 21 symbols for the amino acids and the terminator. The genetic code map is shown in table 1. In the center of the table we list the amino acids. The image of a codon can be obtained locating its first base in the left side of the table, the second in the top and the third in the right side. For example the codon AUC codes for Ile (isoleucine).

The table shows the degeneracy of the code. There are three amino acids, serine (Ser), leucine (Leu) and arginine (Arg) which are coded by six codons. Five of then are quadruplets and two (Ile and Term) are triplets. Mostly of the amino acids are coded by two codons (nine of then) and metionine and tryptophane are the only singlets. The code is universal with few exceptions, this means that all cells in all species uses the same translation rules.

TABLE 1 . The standard genetic code

FIRST base	SECOND base				THIRD base
	U	C	A	G	
U	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.

The AAGC was a consequence of a search for symmetries among the complex Lie algebras to verify if the degeneracy of the genetic code could be obtained by a well-defined mathematical procedure. The idea was to associate the 64 codons to vectors that form a basis in a vector space. This space was require to carry an irreducible representation of a simple complex Lie algebra G. The representation of the algebra G was decomposed as the sum of irreducible representations of a chain of maximal subalgebras $G \supset G_1 \supset \ldots G_n$. At the end of the chain the 64^{th} dimensional irreducible representation of G should break in 21 subspaces with the dimensions given by the degeneracy of the genetic code map. The remaining subspaces that are invariant only under G_n transformations are labeled by the 20 amino acids and by the termination codons. Therefore G_n implements changes in the codons with keeps unchanged the translation provided by the genetic code. This procedure sets the framework for the classification of codons in irreducible representations of the subalgebras that allow us to discuss symm etries and is compatible with several classes of dynamical systems that will ultimately need for a quantitative analysis of the biological content of the model.

The search was made possible due to the Cartan Classification Theorem. It establishes that a simple complex Lie algebra is necessarily a member of one of the four families of classical algebras A_n, B_n, C_n and D_n or it is one of the exceptional algebras G_2, E_6, E_7, E_8 and F_4. The well-known restrictions on the possible dimensions of irreps of a Lie algebra [8] reduces the search for such "codon" representations to a finite number. Once we have selected the list of admissible Lie algebras and the corresponding codon irreps the search continues by exhaustion. It is an easy task for the low ranking algebras but involves a large number of possibilities for the high ranking algebras due to the existence of a large number of chains of subsymmetries. In the search we have used the historic background on this kind of procedure in the field of high energy physics, specifically in the "fever " of the grand-unified theory in the last decades.

Technically the investigation of how an algebra breaks in its maximal subalgebras requires the knowledge of the so called "branching rules" that have been tabulated along this century starting from the seminal papers of Dynkin [9], [10] . A systematization of the procedures used in [1] in the search is underway [11]. The idea is to precise in axiomatic form the principles involved in the search.

The result of the search was the selection of the six-dimensional sympletic algebra sp(6) as the best choice of symmetry. The actual code is obtained by breaking the symmetry in the last step by an operator that is not an invariant of a subalgebra but is a member of a very restricted class of operators in the enveloping algebra of sp(6) [11].

A careful analysis of all the possible chains was made, and there is no perfect symmetry that will generate the genetic code. Nevertheless the sp(6) chain breaking in the sp(4)$\oplus$ su(2) is the one that best reproduces the genetic code degeneracy.

To obtain the genetic code, a ”freezing term“ must be introduced in the last step of the symmetry chain. We correlate the symmetry chain to the evolution of the genetic code and the ”freezing term“ to the ”frozen accident theory“ presented by Crick [12] to explain the stability of the genetic code. If the ”freezing term“ is removed the genetic code generated by this symmetry chain would allow for 26 amino acids and one stop signal. It is interesting that analyzing this problem, with only biological and biochemical considerations, Jukes [13] suggested that if the ”frozen accident“ had not occurred a code of 28 amino acids would have been generated.

The complete symmetry chain proposed for the evolution of the genetic code is:

$$\mathrm{sp}(6) \supset \mathrm{sp}(4) \oplus \mathrm{su}_3(2) \supset \mathrm{su}_1(2) \oplus \mathrm{su}_2(2) \oplus \mathrm{su}_3(2) \supset \mathrm{su}_1(2) \oplus \mathrm{su}_2(1) \oplus \mathrm{su}_3(2)$$

$$\supset \mathrm{su}_1(2) \oplus \mathrm{u}_2(1) \oplus {}^*\mathrm{u}_3(1).$$

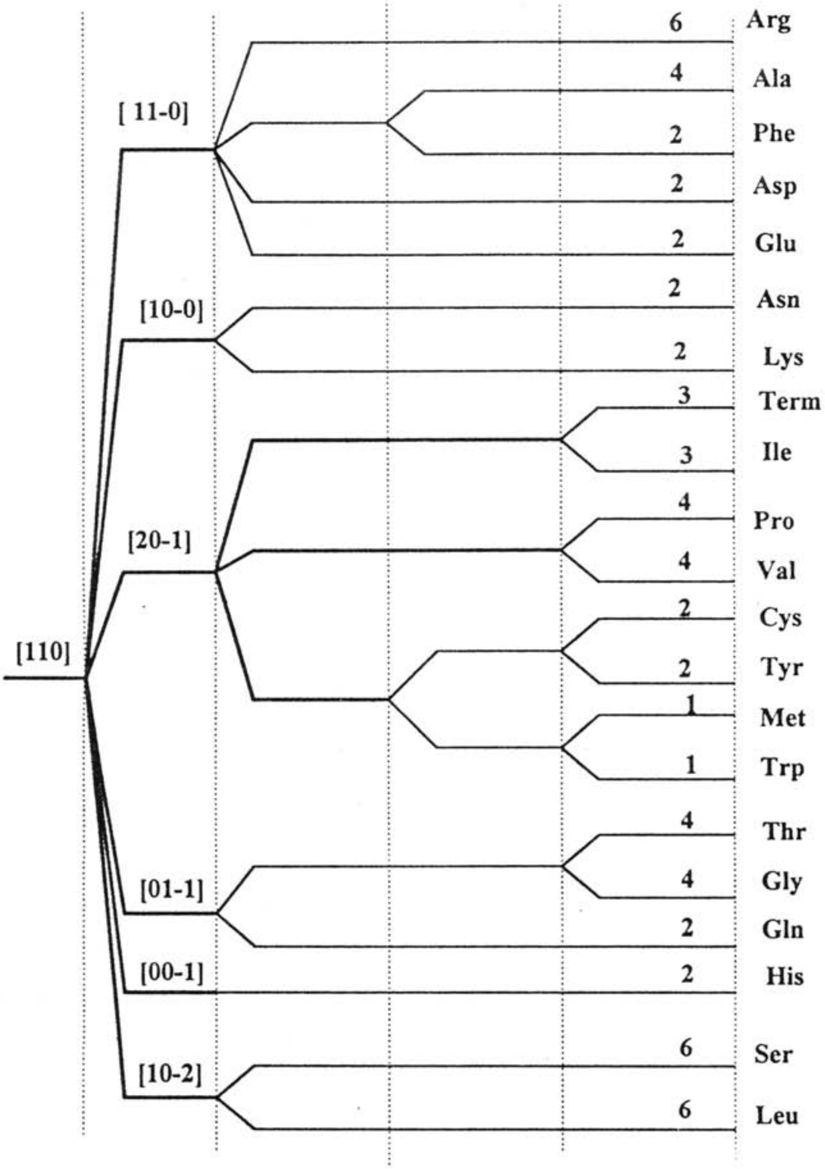
We have four steps in the chain, in the last one, we call ${}^*\mathrm{u}_3(1)$, the star is to indicate that the symmetry breaking is incomplete. Mathematically we can represent this symmetry chain using the Casimir operators of these algebras, in a similar way as it is done in the spectrum generating algebra used in nuclear physics [14]. The operator that furnishes a different eigenvalue for each one of the 21 degenerated states of this symmetry chain, for the (1,1,0) irreducible representation of the sp(6) is:

$$H = h_0 + h_1 \mathcal{L}_4 + q_1 \mathrm{L}_1^2 + q_2 \mathrm{L}_2^2 + q_3 \mathrm{L}_3^2 + p_1 \mathrm{L}_{1\mathrm{Z}}^2 + p_1 (\mathrm{L}_1^2 + \mathrm{L}_2^2)(\mathrm{L}_3^2 - 2)\mathrm{L}_{3\mathrm{Z}}\,,$$

where h_0, q_1, q_2, q_3, p_1 and p_2 are arbitrary constants, $\mathcal{L}_4$ the quadratic invariant operator of sp(4) , $\mathrm{L}_1, \mathrm{L}_2$ and L_3 are the angular momentum operator for the three su(2) involved in the chain. $\mathrm{L}_{1\mathrm{Z}}$ and $\mathrm{L}_{3\mathrm{Z}}$ are the z component of the angular momentum which are invariants of the Abelian subalgebras $\mathrm{u}_1(1)$ and $\mathrm{u}_3(1)$. The term $(\mathrm{L}_1^2 + \mathrm{L}_2^2)(\mathrm{L}_3^2 - 2)$ that multiplies the $\mathrm{L}_{3\mathrm{Z}}$ operator is responsible for the ”freezing“. Schematically, this symmetry chain is shown in figure 1, with the amino acids assignments to each representation.

The canonical chain of Sp(6) occurs when it breaks in the sp(4)⊕su(2) subalgebra leading naturally to the classification of the states in accord to the behavior under su(2) transformations. This algebra admits spinorial and vectorial irreps that are called ”bosonic“ and ”fermionic“ representations, respectively. If we look to this classification from the point of view of the groups and not algebras we find that the vectorial representations are typical of the O(3) group and the spinorial representation belongs to SU(2), the universal covering group. This distinction will show to be important in the biological interpretation of the model.

Figure 1. Representation of the evolution of the genetic code in the symmetry chain sp(6)⊃sp(4)⊕su(2)⊃su(2)⊕su(2)⊕su(2)⊃su(2)⊕u(1)⊕su(2)⊃su(2)⊕u(1)⊕ *u(1). The numbers over the left bars are the dimension of the final representations.



The next step is to assign codons and amino acids to the constructed invariant subspaces. If we have an amino acid, like serine, which has 6 codons it should be attached to a 6-dimensional hyperspace of the irrep. But if we stay only under restriction of the degeneracy we have an incredible high number of possibilities. The case of doublets is an illustrative example: we have 9 of them and consequently we have 9! possibilities of assignment. This problem has been solved recently [15] by combining the tensorial properties of the simpletic algebra pointed above and symmetry principles based in biochemical considerations. A unique assignment has been obtained based in these considerations enforcing the predictability of the model.

The sympletic algebras have not been as exhaustively studied as the unitary and orthogonal similars [16, 17] . They are not so frequent in physical applications and are technically more involved. This means that there is no way to obtain the representation matrices for an arbitrarily irrep of sp(n). In the case of the orthogonal and unitary algebras a general method has been achieved by Gelfand and coworker [18]. Large but unfortunately the results can not be extended to the sympletic case. This means that the representations must be constructed in the specific case and for these several methods are available. The representation matrices for the (110) codon representation of Sp(6) have been constructed by the tensor method [19], by boson operator tecniques [20, 21] and the construction by the Gelfand procedure are under investigation [22]. The results can be used not only to detect symmetry selection rules for the model but also to study the non-standard genetic codes.

Another class of questions is related to the non-standard genetic codes. Until two decades ago, the genetic code was supposed to be universal, i.e valid for all species. With the progress of experiments in which the sequence of RNA and the sequence of amino acids in proteins have been determined in several organisms, non-standard genetic codes have been found. They appear in organelles called mithocondria, but also in more complex organisms and them present small deviations from the standard model. The questions of the relation between the AAGC and the non-standard codes have been preliminary discussed By J.Clella Flores [23] that suggested the possibility to compare the codes in the context of the AAGC. Three other possibilities have been devised recently. The first compares the mitochondrial codes with the same symmetry group advocating that the non-standard codes have been originated from different chains of the same symmetry [24] . The second directs the efforts to compare the codes, with the same group in the same chain, trying to classify the irregularities of the exceptional codes [25]. Finally the general search program done for the standard code must be repeated for all non-standard codes [11].

The most difficult problem involving the AAGC is directly related to analysis. As usual in approach's based in symmetries, the models do not leads automatically to any kind of dynamical system. It only furnishes the general framework for the possible alternatives. Any dynamical systems have to obey the principles of symmetries imposed by the model and have to generate the breaking patterns in each step. This means not only to find a sp(6) invariant system, but a system adapted to the complete chain of subalgebras. A step toward this goal has been already given with the investigation of the invariant polynomials in irreducible representations, by the study of Molien functions [26]. The AAGC fits naturally to a process of spontaneously broken symmetries, analogous to similar process in gauge theories. However several other alternatives other than spontaneously broken symmetries are also possible in the scope of dynamical sy stem theory.

Finally we have to consider the codon-anticodon problem [27]. It is today well established that each codon is recognized by at least one anticodon located in another kind of RNA molecule: the transfer RNA. An incredible fast process in the determination of the genoma of several species is today in course and certainly will produce a rich set of new experimental information. The relation between codons and anticodons in the model remain an open question.

We conclude our brief outline of the open problems in the AAGC pointing that, it is not possible today to accept the alternative that the Sp(6) symmetry could be "a remarkable coincidence“ as have been considered in [1]. The successive agreement between the AAGC and the biological considerations are strongly in favor to the interpretation of the sympletic symmetry as the dominant symmetries in the genetic code. The model remains a rich tool for the investigation of the regularities in the genetic code and its structure is far from been exhausted or even fully understood.

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] J.E.M. Hornos and Y.M.M. Hornos, A Search for Symmetries in the Genetic Code, J.Biol.Phys. **20** (1994) 289-294.

[3] J.Maddox, The Genetic Code by Numbers, Nature **367** (1994) 111.

[4] I.Stewart, Broken Symmetry in the Genetic Code ?, New Scientist (5 March 1994) 16.

[5] M.Forger, Symmetry Breaking in the Genetic Code, 41º Seminário Brasileiro de Análise (1995).

[6] Y.M.M. Hornos, M. Forger, J.E.M. Hornos, The Algebraic Approach to the Genetic Code, (in preparation).

[7] F.H. Crick, L. Barnett, S. Brenner and R.J. Watts-Tobin, General Nature of the Genetic Code for Proteins, Nature **192** (1961), 1227-1232.

[8] B.G. Wybourne, Classical Groups for Physicists

[9] E.B. Dynkin, Maximal Subgroups of the Classical Groups, Trudy Mosk. Mat. O-va., **1**, 39 (1952). Transl. in Am. Math. Soc. Transl. (2), **6**,245 (1965).

[10] E.B. Dynkin, The Structure of Semisimple Algebras, Usp. Mat. Nauk (N.S.), **2**, 59 (1947). Transl. in Am. Math. Soc. Transl. (1), **9**,308 (1962).

[11] L.Braggion, A Procura de Simetrias no Código Genético, dissertação de mestrado (IFSC-USP),em preparação.

[12] F.H. Crick, The Origin of the Genetic Code, J.Mol.Biol. **38** (1968)367-379.

[13] T.H. Jukes, Evolution of the Amino Acid Code: Inferences from Mitochondrial Codes. J.Mol.Evol. **22** (1983) 219-225.

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

[15] J.E.M. Hornos, Y.M.M. Hornos and M.Forger, Amino Acids and Codon Assignments in the Algebraic Approach to the Genetic Code, in preparation.

[16] M.Cerkaski, Branching Rules for sp(2N) Algebra Reduction on the Chain sp(2N-2)xsp(2),J.Math. Phys. **28** (1987) 989-990.

[17] M.D. Gould, Representation Theory of the Sympletic Groups.I, J.Math.Phys. **30** (1989) 1205-1218.

[18] I.M. Gelfand and M.L. Zetlin, Finite Dimensional Representations of a Group of Unimodular Matrices, Dokl. Akad. Nauk SSSR **71** (1950) 825-828.

[19] M. Forger, E.Bernardes and J.E.M.Hornos, Tensorial Construction for the Codon Representation of Sp(6).

[20] E.Chacon and M. Moshinsky, Bases of Irreps of the Chain of Lie Algebras $sp(6)\supset sp(4)\oplus su(2)\supset su(2)\oplus su(2)\oplus su(2)$. (private communication).

[21] M. Barbosa, Construção das matrizes da representação (110) do Sp(6) através de operadores bosônicos. Dissertação de Mestrado (IFSC-USP).

[22] E.S. Bernardes, Analitic Matrix Elements for sp(4,C) Lie Algebra, in preparation.

[23] J. Chela-Flores, Some Physical Problems in Biology, J.Biol.Phys **20** (1994).

[24] Y.M.M. Hornos, Symmetry in the Evolution of the Mitochondrial Genetic Code, to be presented in the XXI Int. Coll on Group Theo. Meth. in Phys.- Goslar (1996)

[25] Y.M.M. Hornos, M.Forger and J.E.M. Hornos, Analysis of the Deviations from the Standard Genetic Code in the Algebraic Approach (in preparation).

[26] M.Forger, Invariant Polynomials and Molien Functions, Relatório Técnico RT-MAP-9503 , IME-USP

[27] S. Ozawa, T.H. Jukes, K. Watanabe and A. Muto, Recent Evidence for the Evolution of the Genetic Code, Microbiological Reviews, **56** (1992), 229-264.

José Eduardo Hornos
Departamento de Física e Ciência dos Materias
Instituto de Física de São Carlos
Universidade de São Paulo
Caixa Postal 369,13560 - São Carlos, S. P.
Brasil

Yvone M. M. Hornos
Sapra - Serviço de Assessoria e Proteção Radiológica S/C Ltda
Rua: Dr. Orlando Damiano, 2160, São Carlos, S. P., 1356-990
Brasil