IC/91/17

INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS

A SEARCH FOR SYMMETRIES IN THE GENETIC CODE

José Eduardo M. Hornos

and

Yvone M.M. Hornos

INTERNATIONAL ATOMIC ENERGY AGENCY

UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION

1991 MIRAMARE - TRIESTE

IC/91/17

International Atomic Energy Agency
and
United Nations Educational Scientific and Cultural Organization
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS

A SEARCH FOR SYMMETRIES IN THE GENETIC CODE *

José Eduardo M. Hornos ** and Yvone M.M. Hornos **
International Centre for Theoretical Physics, Trieste, Italy.

ABSTRACT

A search for symmetries based on the classification theorem of Cartan for 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, universal 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.

MIRAMARE – TRIESTE
January 1991

* To be submitted for publication.

** Permanent address: 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, SP, Brazil.

1. INTRODUCTION

The genetic code for the construction of protein molecules is composed of triplets of four fundamental bases: adenine, guanine, cytosine and uracil or thymine. Each one of the amino acids is coded by one or more triplets: the code is therefore degenerate. When the relation between the codons and the amino acids were established it was clear that a systematic investigation should be performed in order to understand the code.

A first attempt to explain the relation between the amino acids and the sequence of the bases was the proposal of a stereochemical theory (Woese *et al.*, 1966a; 1966b). This theory claims the existence of physico–chemical interactions between the bases in the triplets and the amino acids which "fit stereochemically" with the particular combination.

Another point of view was propsed by Crick in the so–called Frozen Accident Theory (Crick, 1968). In its extreme form the Frozen Accident Theory implies that the allocation of codons to the amino acids at this point was entirely a matter of chance. In the same spirit Jukes 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. The irregularities of the code are a testimony to the freeze (Jukes, 1983).

A different model was proposed under the name of the Co–Evolution Theory (Wong, 1975; 1976; 1980; 1983). In this theory a few primordial amino acids 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 physico–chemical properties of the amino acids and the code itself. Correlations between polarity, hydrophobicity, etc. and the codon assignments have been the subject of extensive investigations by many authors (Di Giulio, 1989; Zimmerman *et al.*, 1968, Weber and Lacey, 1978).

The main goals of the present paper are not only to report the search for the "best" underlying symmetry of the universal code, but also to show that the theory of symmetries can provide a mathematical framework in which a dynamical evolutionary pattern for the code emerges naturally. Also, rules for branching of the primitive amino acids into new ones are especially harmonious using this method.

From the result of this search a new class of models for the evolution of the genetic code based on the notion of symmetries has emerged. The group theory provides a set of well–defined simple rules for the construction of only a few "symmetry codes" among an enormous number of possibilities allowed by combinatorial analysis. Most of them give "non–sense" genetic codes, but only one based on the simplectic symmetry is appropriate for genetics. In this context it should be stressed that the appearance of only left–handed amino acids in natural proteins organisms may be a manifestation of a symmetry breaking in the evolution of life on earth (Salam, 1990).

The ideas of symmetries have been extensively used in the description of inanimate matter. The quark model (Gell–Mann, 1962) *) for the strong interacting subnuclear particles and the electroweak theory (Glashow, 1961; Weinberg, 1967; Salam, 1968) are striking examples in which symmetries are the underlying organizational principle for the very existence and correlations of the fundamental particles. From the subnuclear particles to nuclei we also find that symmetry groups are a useful tool in the description of energy levels and degeneracy of the spectrum of nuclei when they are analyzed by the Interactive Boson Model (IBM) (Arima and Iachello, 1987). A similar approach, the vibron model, has been successfully used in the analyses of molecular spectra (Iachello and Levine, 1982; Homos and Iachello, 1989).

Symmetry is easily encountered in the geometric structure of biological molecules. The use of discrete finite groups is an important tool in the determination of the crystallographic structure of proteins. Geometric symmetries are also present in more complex systems such as viruses. The icosahedric shape and presence of five–fold symmetry are common features of viruses.

A typical continuous symmetry is the rotation invariance in classical and quantum mechanics. These groups are appropriate for the description of dynamical processes. They provide a set of selection rules and conservation theorems. The richness of these symmetries led the authors to begin the search with the continuous groups.

This paper is organized as follows. In Section 2 we present a review of the properties of the genetic code introducing the basic ideas and the notation. In Section 3 the basic notions related to the Lie algebra and the Cartan classification are established. Instead of a more formal approach we focus our attention on the fundamental ideas and on the use of tables that allow the modelling without an extensive knowledge of group theory and Lie algebras. In Section 4 a sample model is presented. The theoretical model should be considered as an example rather than a complete theory. As shown in that section a more realistic model requires a very careful analysis of the phenomenological data which is beyond the scope of this paper. Section 5 is reserved for the conclusions.

2. THE GENETIC CODE

The genetic code is a triplet code, consisting of 64 triplets of the four bases adenine (A), guanine (G), cystosine (C) and uracil (U) whose universality have been shown in most prokaryotic and eukaryotic species studied. Since the monumental achievement of the cracking of the code, questions such as: was the code initially as it is or did the code evolve to the point it is today? does the nature of the code manifest itself in the physico–chemical properties of the 20 amino acids? have been the subject of many studies including this one. Even before the universal code presented in Table 1 was totally known, a number of regularities had been pointed out. For example:

*) For a technical review: *An Introduction to Quarks and Partons* by F.E. Close, Academic Press.

1. the 20 amino acids are not distributed at random among the 64 triplets,
2. if the third base is a uracil (cytosine) it can be interchanged by a cytosine (uracil) not changing the amino acid it codes,
3. in most cases an adenine (guanine) in the third base can be replaced by a guanine (adenine) also not changing the amino acid it codes, the exceptions to this rule are methionine and tryptophan that are coded by only a single codon each,
4. in half of the cases, 8 out of 16, the pair of the first two bases (XY) represents a single amino acid, implying that the third base can be any of the possible four,
5. codons representing a single amino acid start with the same pair of bases except for the amino acids that have multiplicities of six codons, being arginine (ARG), leucine (LEU) and serine (SER). Other such regularities that have been noted include the fact that, all codons with U in the second position code for hydrophobic amino acids. We will not list all these observations here since our objective is to give a precise way to "derive" all these regularities. We only stress that the third place of the triplet is clearly more redundant than the combination of the first two.

Let us turn now our attention to the universal code shown in Table 1. There, a three letter code is given for each of the 20 fundamental amino acids plus the termination code and A, C, U, G represent the four bases. An alternative to these lengthy names is to describe the bases by vectors

$$U = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \; ; \quad G = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \; ; \quad C = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \; ; \quad A = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} . \tag{1}$$

The notion of symmetry emerges when we allow a matrix to act on these vectors via matrix multiplication. A useful example is given by the matrices L_+, L_- and L_z where,

$$L_+ = \begin{bmatrix} 0 & \sqrt{3} & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & \sqrt{3} \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad L_- = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \sqrt{3} & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & \sqrt{3} & 0 \end{bmatrix} \quad L_z = \frac{1}{2} \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix} \tag{2}$$

$$L_z \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \frac{3}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \; ; \quad L_z \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$L_z \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \; ; \quad L_z \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = -\frac{3}{2} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} . \tag{3}$$

Each chemical base can be identified by an half integer $U = \frac{3}{2}, G = \frac{1}{2}, C = -\frac{1}{2}$ and $A = -\frac{3}{2}$. The other two operators, L_+ and L_-, are called raising and lowering operators because they change the base to another one

$$L_+ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \sqrt{3} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \; ; \quad L_+ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$L_+ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \sqrt{3} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \; ; \quad L_+ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = 0 \tag{4}$$

and

$$L_- \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = 0 \; ; \quad L_- \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \sqrt{3} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$$L_- \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \; ; \quad L_- \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \sqrt{3} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{5}$$

The operator L_+ changes the state identified with cytosine to adenine adding a multiplicity factor. Adenine will transform to guanine and guanine to uracil. The last action annihilates the base. The operator L_- also changes the base in the reversed order. These properties have not been used in this paper but may be promising in the study of genetic mutations.

The important properties of these matrices is that they obey the following relations that can be easily verified.

$$\begin{aligned} L_z \cdot L_+ - L_+ \cdot L_z &= L_+ \\ L_z \cdot L_- - L_- \cdot L_z &= L_- \\ L_+ \cdot L_- - L_- \cdot L_+ &= 2 L_z \end{aligned} \tag{6}$$

A fourth important matrix in the mathematical theory of symmetries is formed by the combination of:

$$\left(\frac{1}{2}\right)(L_+ \cdot L_- + L_- \cdot L_+) + L_z^2 = L^2$$

it is proportional to the identity matrix with the proportionally coefficient $\frac{15}{2} = \left(\frac{3}{2}\right) \times \left(\frac{3}{2} + 1\right)$.

We summarize our group theoretical description of the fundamental bases.

1. Each base is identified by a vector in a four dimensional "representation" space.
2. The matrices L_z, L_+ and L_- are technically generators of a representation of the rotation group $SU(2)$. They relate the bases amongst themselves.
3. The operator L^2 does not distinguish the bases, it gives $\ell(\ell+1)$ times each base $\left(\ell = \frac{3}{2}\right)$.

4. The matrix L_z selects each base associating with it one of the half integers $\frac{3}{2}$, $\frac{1}{2}$, $-\frac{1}{2}$ or $-\frac{3}{2}$. The matrix L_z generates a single unidimensional space with the simplest $SO(2)$ algebra related to the well–known crystallographic point groups.

In Table 2 we summarize our notation.

3. THE CARTAN CLASSIFICATION THEOREM

A straightforward generalization of Eq.(6) defines a general Lie algebra as a set of large matrices obeying the equations

$$[H_i, F_\alpha] = \lambda_{i\alpha} E_\alpha \quad \alpha = \pm 1, \pm 2, \ldots \pm N \tag{8a}$$

$$[E_\alpha, E_{-\alpha}] = \Sigma \; \rho_i^\alpha \; H_i \tag{8b}$$

$$[H_i, H_j] = 0 \quad i, j = 1, \ldots r \tag{8c}$$

$$[E_\alpha, E_\beta] = N_{\alpha\beta} E_{(\alpha+\beta)} \; . \tag{8d}$$

The more complex algebras have r commuting matrices like the L_z generator of the rotation group and also more than one pair of raising and lowering operators. Eqs.(8) are generalizations of Eqs.(6). Eq.(8c) is the generalization of the trivial equation $[L_z, L_z] = 0$ for the $SU(2)$ group and defines the so–called Cartan commuting sub–algebra. Finally Eq.(8d) expresses the closure of the algebra. The number of matrices defines the order of the group and the number of commuting matrices gives the rank of the algebra.

The Cartan classification theorem states that there are only nine different families of Lie algebras labelled A_n, B_n, C_n, D_n, E_6, E_7, E_8, F_4 and G_2. The order, the Cartan label and a second label generally used in applications are presented in Table 3. The rotation algebra $SU(2)$ is the simplest one. The point to stress here is that the problem of how many and which algebras are possible has been definitely solved by Cartan. The handling of these algebras, although tedious, is a relatively straightforward procedure. Recently, computer codes have become available that perform the analysis of the algebra.

Another major achievement in this area was the corresponding classification of all representations of the compact simple algebras. The representations of the algebras are labelled by r ordered integers. The representations of the larger algebras contain the smaller ones thereby forming chains. The dimension of all representation algebras of rank smaller than 10 are tabulated (McKay and Patera, 1981). The branching rules to form chains are also present in the same table. The situation is totally analogous to the tabulation of the representations of finite groups commonly used in the spectroscopic analysis of molecular structures in all branches of chemistry and biochemistry.

In particle physics the use of symmetries has been of enormous importance for establishing the quark model, which not only explained the presence of the observed particles but also predicted particles, that were subsequently observed.

4. A MODEL FOR THE CODE

Using combinatorial analysis it is estimated that at least 10^{71} to 10^{84} different genetic codes like our contemporary one are possible (Bertman and Jungck, 1978), i.e., by starting with the 64 codons and arranging different ways of distributing them among the 20 amino acids and one termination code. 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 the chains of subgroups. Since we need to start with 64 codons, we look for groups that have a 64 dimension representation and, the inspection of the tables of dimension leave us with only 8 possible algebras, listed in Table 4.

At this point, there are 64 codons, but no mathematical distinction between them exists. In order to make such a distinction, we look for all possible chains of each of the 8 algebras. An algebra of rank r will only break in algebras of rank smaller or equal to r. Table 5 lists the largest possible algebras in the chain of each one of the algebras of Table 4. Following the "rules" given in Table 5 only 18 different chains up to $Sp(6)$ are possible and all of them may stop at any point. At this point a search for the chain whose multiplicity of the non–degenerated states reproduces the multiplicity attributed to each one of the 20 amino acids and the termination signal in the universal code was made.

Since the genetic code contains two singlets, nine doublets, two triplets, five quadruplets and three sextuplets, all representations with dimension larger than six will have to be broken. Also the chains of all the Cartan algebras terminate in $SU(2)$ which has the properties already described in Section 2.

Careful analysis of all the possible chains was made, including the $SO(13)$ and $SO(14)$, and there is no perfect symmetry conservation, nevertheless the $Sp(6)$ chain $Sp(4) \otimes SU(2)$ *is the one* that best reproduces the genetic code. This chain shows a "quenching" at the last step, i.e., the symmetry lost occurred at a point that can be detected very precisely at the last step of the evolution. It is interesting to emphasize here that the $Sp(n)$ groups are highly non commutative algebras, and the genetic code is also a non commutative code, i.e., while ACU codes Phe, CAU codes His, and this is true for the entirety of the code. So the fact that the starting group is $Sp(6)$ and the chain leads via $Sp(4) \otimes SU(2)$ to the $SU(2) \otimes SU(2) \otimes SU(2)$ symmetry can be attributed to the non–commutability of the code.

The genetic code approximate symmetry chain and the perfect symmetry are displayed respectively in Figs.1(a) and (b), where each state is indicated by a horizontal bar with the multiplicity specified by the number over the bar; the horizontal axis indicates the evolution;

Step 1: the $Sp(6)[1,1,0]$ representation evolved to 6 different representations of $Sp(4) \otimes SU(2)$, breaking the degeneracy of 64 into six different states with multiplicity 16, 4, 20, 10, 12 and 2, these states correspond to the primordial amino acids;

Step 2: the $Sp(4) \otimes SU(2)$ breaks into $SU(2) \otimes SU(2) \otimes SU(2)$, here each one of the $Sp(4)$ representation subdivides into representations of $SU(2) \otimes SU(2)$, the state with degeneracy $16 \rightarrow 16+6+2+2$; the state with degeneracy $20 \rightarrow 6+8+6$, and the others are indicated in the figure;

Step 3: the second $SU(2) \rightarrow \tilde{0}(2)$ with $L^2 z$, again the multiplicities that changed from the one in Step 2 are indicated;

Step 4: the last $SU(2) \rightarrow 0(2)$ as L_z it was at this stage of evolution that the freeze occurred; the ramification of the state that started in Step 1 with multiplicity 16, 4 and 20 suffered all the chain breaking, however the states with multiplicity 10, 12 and 2 only the $Sp(4) \otimes SU(2) \supset SU(2) \otimes SU(2) SU(2) \supset SU(2) \otimes \tilde{0}(2) \otimes SU(2)$ occurred and finally the state at Step 1 with multiplicity 10 suffered only partially the total symmetry breaking, in Steps 2 and 3 the multiplicity of $10 \rightarrow 8+2$ and in Step 4 the $8 \rightarrow 4+4$ and the state with degeneracy 2 would normally under the action at $SU(2) \rightarrow 0(2)$ subdivide in $1+1$ but to reproduce the multiplicity of the genetic code this symmetry breaking must stop;

Step 5: in Fig.1(b) we show how the code would be if the symmetry breaking was allowed to continue. At this stage a perfect symmetry code could be written.

Fig.2 shows two more possibilities of chains that up to the last step look very promising, however in these cases, there is no way to match the genetic code with a quenching or change of the symmetry in a regular way.

Let us turn our attention back to Fig.1. Here we have first matched the multiplicities of the code, but there is a large ambiguity in the assignment of the doublets, quadruplets and sextuplets. The analysis for the tentative assignment shown in Fig.1 was based on two hypotheses. We first assume that since Asp, Ser, Ala and Gly are the most easily synthesized amino acids in laboratory experiments designed to reproduce the original soup, they must be the primordial amino acids when possible and second, that the existence of a termination code is primordial also. If this is assumed we see from the structure of the branching that there is no mixing after the first break, but only that some of the codons that in one step of evolution corresponded to a particular amino acid will in the next step correspond to a different one with a parental relationship. In the model, $Sp(4) \otimes SU(2)$ was the primordial step at which point there was 6 primordial amino acids with a multiplicity of codons of 20, 16, 12, 10, 4 and 2. The assignment shown in Fig.1 was made according to the calculated polarities and the assignment of primordial amino–acids:

1. if Step 4 corresponds to the present code and the termination signal Term was present in the primordial soup and as only the state with multiplcity 20 will break giving triplets the 20– fold state in Step 1 must be assigned to Term;

2. if Asp was one of the original amino acids, and there are only two negatively charged amino acids at neutral pH and if only one is taken to be primordial we can assign both of them, Asp and Glu, in the same multiplet the one with multiplicity 16 at Step 1;

3. the state with multiplicity 12 subdivides at Step 2 into two sextuplets and does not subdivide further, thus Arg and Ser, or Arg and Leu, or Ser and Leu must be assigned here. We allocate Ser as primordial and Leu as a subdivision of Ser.

4. Since Ser and Leu are already assigned, Arg must be in the subdivision of the state with multiplicity 16. However Leu is not considered a primordial amino acid but rather Asp, which requires a state with multiplicity four in Step 4, is considered primordial, other polar amino acids such as Pro and His can be allocated with it in this multiplet.

5. Two states are left, respectively with multiplicity 10 and 2, Val and Thr must be allocated in the first because both of them requires in Step 4 a multiplicity of four.

6. Finally the remaining amino acids are allocated in accordance with amino acids polarity.

At Step 2 the $Sp(4) \otimes SU(2)$ chain goes to $SU(2) \otimes SU(2) \otimes SU(2)$, and at this point the model resembles the product of three $SU(2)$ groups, one for each base of the codon, the non–commuting properties of the code is preserved due to the $Sp(4)$ symmetry. At this step the primordial amino acids subdivide themselves into 14 amino acids with only one of them Cys, remaining inalterated. We note that Jukes suggests an archetypal code containing 14 or 15 amino acids (Jukes, 1983). Again the symmetry is broken in the second $SU(2)$, not going to $0(2)$ and L_z (as explained in Section 2) but as the square of L_z. We denote this type of symmetry by $\bar{0}(2)$ and after the break a code with 16 amino acids results. The evolution of the code proceeds in Step 4, again breaking the last $SU(2)$ in $0(2)$ (L_z). 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 $0(2)$, however they did not subdivide and the code was frozen with only 20 amino acids, the "symmetry perfect code" would have 27 amino acids. Again Jukes suggested that the code should generate 28 amino acids if the freeze had not occurred.

The mathematical framework of these continuous groups are essentially given by combination of group operators that acting on their representations furnishes eigenvalues that will be associated to physical quantities. The operator that reproduces the multiplicities of the code is:

$$\mathcal{H} = h_0 + h_1 \mathcal{L}_4 + q_1 L_1^2 + q_2 L_2^2 + q_3 L_3^2 + p_1 L_{z_2}^2 + p_2 (L_1^2 + L_2^2)(D_3 - 3) L_{z_3} \qquad (9)$$

where $h_0, h_1, q_1, q_2 p_1$ and p_2 are arbitrary constant, $\mathcal{L}_4$ 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 component of the angular momentum operator of the $0(2)$; 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. $|N_1, N_2, K_1, K_2, K_3, m_1, m_2, m_3 >$, where N_1, N_2 corresponds 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 $0(2)$ is given by

$$h = h_0 + h_1 [(N_1 + N_2)(N_1 + N_2 + 4) + N_2(N_2 + 2)] + \sum_{i=1}^{3} q_i (1/4)(K_i)(K_i + 2) + P_1 m_2 + P_2 (1/4)(K_1^2 + K_2^2)(K_3 - 2) m_3 . \qquad (10)$$

In order to make these ideas more clear we constructed a first example of how this mathematical framework can be used to calculate quantities related to the genetic code. Using experimental polarities (Grantham, 1974) we minimize Eq.(10) to find the values for the arbitrary constants, that gives the minimum square deviation between the experimental and calculated polarities, the final result also guided the assignment of the amino acids with same multiplicity to the different group theoretical states.

In Table 6 we show the group theoretical assigment for the amino acids, the values used for the arbitrary constants and the calculated and experimental values of the polarities for the 20 amino acids, Fig.3 shows a bar graph of our results. 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 all the physico–chemical properties of the amino acids, 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 physico–chemical property that guided the evolution of the genetic code in the symmetry path.

The assignment of the group theoretical states to the amino acids can also be made including arguments related to the biopathways of the amino acids, this must be reflected in the connectivity given by the specific chain followed by evolution.

5. CONCLUSIONS AND SUMMARY

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. An example of the flexibility of this theory is the codon assignment, in the group theoretical structure, several biological possibilities can be used in the assignment of the states with defined multiplicity, this freedom is an advantage of the symmetry principles where the imposed restrictions are very general and we can accommodate other restrictions coming from a different aspect of the problem in analysis. Group theoretical principles will not tell which of the amino acids were primordial, but will only answer questions relating the multiplicities and connectivities between different states (each state refer to one amino acids) with a defined multiplicity. Also the non–universality of the code can be analyzed in this formalism because different forms at breaking the symmetry will lead to different codes, our programme includes the study of other codes and we will search for the regularities among the model.

Questions such as why should SER be six–fold degenerate and VAL only four–fold, cannot be answered by this model, but prediction such as (1) there were two six–fold correlated amino acids in an earlier stage of evolution, (2) that the differentiation of TRP and MET occurred only in the last stage of evolution, (3) there is a two–fold amino acid (in our model CYS) that has not been modified since the beginning, etc. are examples of predictions that can be directly derived from Fig.1.

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 on the $Sp(4) \times SU(2)$ symmetry have emerged, in general lines we establish that:

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 $Sp(4) \otimes SU(2) \supset SU(2) \otimes SU(2) \otimes SU(2)$ 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 amino acids with different multiplicities can be made.

The mathematical description of the amino acids, in terms of a set of numbers, presented here and used to calculate the Granthan polarities, immediately suggests several topics for further investigations, for example, the full set of amino acids properties (Nakai, Kidera and Kanehisa (1988) – the database was kindly furnished to us by Peter Sibbald, EMBL) must be be analysed with this new model. A more ambitious project is to correlate the numbers that define each amino acid with important protein properties such as the folding. The study of different codes, e.g.; the mammalian mitochondrias, based in symmetry theory can be used to establish an hierarchy between the different codes in the evolution process.

Finally the newly found fact that there are only two different classes of synthases for the 20 amino acids must also be studied by this approach.

Acknowledgments

The authors wish to thank the hospitality and useful suggestions of several institutions and scientists during the preparation of this paper. They are: at the Center for Theoretical Physics, Yale University, Professors S.W. Mac Dowell and G. Gursey; at UFPe (Brazil), Professor R. Ferreira; at USP (Brazil), Professor Y.P. Mascarenhas; at Institut de Biologie Moleculaire e Celulaire, (Strasbourg), R. Gigier and, finally, at the International Centre for Theoretical Physics, Professors J. Chela Flores and Abdus Salam. Finally, the authors would like to thank the International Atomic Energy Agency and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste. One of the authors (Y.M.M.H.) was supported by Conselho Nacional de Pesquisas (CNPq).

REFERENCES

Arima A. and Iachello F. (1987). The Interacting Boson Model, *Monographs on Mathematical Physics*. Cambridge: Cambridge University Press.

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

Crick F.H.C. (1968). J. Mol. Biol. **38**, 376.

Di Giulio M. (1989). J. Mol. Evol. **29**, 191.

Gell–Mann M. (1962). Phys. Rev. **125**, 1067

Glashow S.L. (1961). Nucl. Phys. **22**, 579.

Grantham R. (1974). Science **185**, 862.

Iachello F. (1981). Chem. Phys. Lett. **78**, 581.

Iachello F. and Levine R.D. (1982). J. Chem. Phys. **77**, 3046.

Hornos J.E. and Iachello F. (1989). J. Chem. Phys. **90**, 5284.

Jukes T.H. (1983). J. Mol. Evol. **19**, 219.

Jungck J. (1982). J. Mol. Evol. **11**, 211.

McKay W. and Patera J. (1981). *Tables of Dimension, Indices, and Branching Rules for Representations of Simple Lie Algebras*. New York: Marcel Dekker.

Miyata T., Miyzawa S. and Yasunga T. (1979). J. Mol. Evol. **12**, 219.

Nakai K., Kidera A. and Kanehisa M. (1988). Prot. Eng. **2**, 93.

Orgel L.E. (1968). J. Mol. Evol. **38**, 381.

Salam A. (1968). Proc. 8th Nobel Symposium, p.367, Aspenasgarden Almquist and Wilksell.

Salam A. (1990). "The role of chirality in the origin of life", ICTP, Trieste, preprint No.IC/90/277.

Weber A.L. and Lacey J.C. Jr. (1978). J. Mol. Evol. **11**, 199.

Weinberg S. (1967). Phys. Rev. Lett. **19**, 1294.

Woese C.R., Dugre D.H., Dugre S.A., Kondo M. and Saxinger W.C. (1966a). Cold Spring Harbor Symp. Quant. Biol. **31**, 723.

Woese C.R., Dugre D.H., Saxinger W.C. and Dugre S.A. (1966b). Proc. Natl. Acad. Sci. USA **55**, 966.

Wong J.T.F. (1975). Proc. Natl. Acad. Sci. USA **72**, 1909.

Wong J.T.F. (1976). Proc. Natl. Acad. Sci. USA **73**, 2336.

Wong J.T.F. (1980). Proc. Natl. Acad. Sci. USA **77**, 1083.

Wong J.T.F. (1983). Proc. Natl. Acad. Sci. USA **80**, 6303.

Zimmerman J.M., Eliezer N. and Simba R. (1968). J. Theor. Biol. **21**, 170.

TABLE CAPTIONS

Table 1 The genetic code.

Table 2 Alternative descriptions of the four bases, using the $SU(2)$ algebra in its four dimensional representation. The first row shows the chemical formula of the four bases; the second throw the "new" representations, and the bottom row the $SU(2)$ chain with the labels ℓ and m and the relation between them.

Table 3 The Cartan classification of the compact algebras.

Table 4 Algebras of rank less than 10, that have a sixty–four dimension representation.

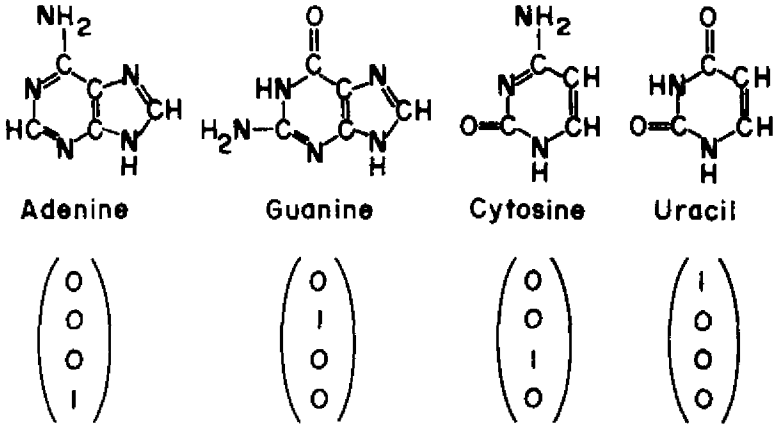
Table 5 Maximal decomposition of the Cartan groups with representation in 64 dimension.

Table 6 Group Theoretical assignment of the amino acids and polarities calculate by the present model.

Table 1

First position	Second position				Third position
	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

Table 2



$SU(2) \supset O(2)$

ℓ $\quad$ m , $m = \ell, \ell-1, \ldots, -\ell$

$\ell = 3/2$, $m = 3/2, 1/2, -1/2, -3/2$

$(\ell, m) = (3/2, 3/2); (3/2, 1/2)$; $(3/2, -1/2)$; $(3/2, -3/2)$

Table 3

CARTAN LABEL	GROUP LABEL	ORDER
A_ℓ	$SU(\ell+1)$	$\ell(\ell+2)$
B_ℓ	$SO(2\ell+1)$	$\ell(2\ell+1)$
C_ℓ	$Sp(2\ell)$	$\ell(2\ell+1)$
D_ℓ	$SO(2\ell)$	$\ell(2\ell+1)$
G_2	G_2	14
F_4	F_4	52
E_6	E_6	78
E_7	E_7	133
E_8	E_8	248

Table 4

CARTAN LABEL	NAME	RANK	REPRESENTATION
A1	SU(2)	1	[63]
A2	SU(3)	2	[3,3]
A3	SU(4)	3	[1,1,1]
C2	Sp(4)	2	[3,1]
C3	Sp(6)	3	[1,1,0]
B6	SO(13)	6	[0,0,0,0,0,1]
D7	SO(14)	7	[0,0,0,0,0,0,1]
G2	G2	2	[1,1]

Table 5

CARTAN GROUP	RANK	MAXIMAL SUBGROUPS
SU(2)	1	O(2)
SU(3)	2	SU(2), SU(2)
Sp(4)	2	SU(2), SU(2)●SU(2)
G2	2	SU(2), SU(2)●SU(2), SU(3)
SU(4)	3	SU(2), SU(2)●SU(2), Sp(4)
Sp(6)	3	SU(2), SU(2)●SU(2), Sp(4)●SU(2), SU(3)
SO(13)	6	SU(2), SU(2)●SU(2)●SO(9), SO(12), SU(2)●SO(10), SU(4)●SO(7), Sp(4)●SO(8)
SO(14)	7	SU(7), SO(13), Sp(6), Sp(4), G2, SU(2)●SU(2)●SO(10), SU(4)●SO(8), SU(2)●SO(11), Sp(4)●SO(9), SO(7)●SO(7)

Name	Group Theoretical State $N_1,N_2;L_1,L_2,L_3;m_1,m_2,m_3$	Polarities Calc.	Exp. (*)
Singlets			
Trp	$\|2,0;0,2,1;0,0,+1/2\rangle$	7.772	5.3
Met	$\|2,0;0,2,1;0,0,-1/2\rangle$	6.592	5.2
Dublets			
Cys	$\|0,0;0,0,1;0,0,\pm 1/2\rangle$	3.902	4.8
Lys	$\|0,1;0,0,1;0,0,\pm 1/2\rangle$	9.022	10.1
Asn	$\|2,0;0,2,1;0,\pm 1,1/2\rangle$	8.772	10.0
Gln	$\|2,0;0,2,1;0,\pm 1,-1/2\rangle$	7.592	8.6
Tyr	$\|1,0;0,1,0;0,\pm 1/2,0\rangle$	5.680	5.4
Phe	$\|1,0;1,0,0;\pm 1/2,0,0\rangle$	5.055	5.0
Asp	$\|1,1;0,1,0;0,\pm 1/2,0\rangle$	12.080	13.0
Glu	$\|1,1;1,0,0;\pm 1/2,0,0\rangle$	11.455	12.5
His	$\|1,1;1,2,0;\pm 1/2,0,0\rangle$	7.055	8.4
Triplets			
Ile	$\|2,0;2,0,1;(\pm 1,0),0,1/2\rangle$	5.592	4.9
Term	$\|2,0;2,0,1;(\pm 1,0),0,-1/2\rangle$	6.772	--
Quadruplets			
Val	$\|0,1;1,1,1;\pm 1/2,\pm 1/2,-1/2\rangle$	5.302	5.6
Thr	$\|0,1;1,1,1;\pm 1/2,\pm 1/2,+1/2\rangle$	5.892	6.6
Gly	$\|2,0;1,1,1;\pm 1/2,\pm 1/2,-1/2\rangle$	8.452	7.9
Pro	$\|2,0;1,1,1;\pm 1/2,\pm 1/2,+1/2\rangle$	7.862	6.6
Ala	$\|1,1;1,2,0;\pm 1/2,\pm 1,0\rangle$	8.055	7.0
Sextuplets			
Leu	$\|1,0;0,1,2;0,\pm 1/2,(\pm 1,0)\rangle$	5.115	4.9
Ser	$\|1,0;1,0,2;\pm 1/2,0,(\pm 1,0)\rangle$	5.740	7.5
Arg	$\|1,1;2,1,0;(\pm 1,0),\pm 1/2,0\rangle$	6.679	9.1

Parameters

$h_0 = 3.88$; $h_1 = 0.64$; $q_1 = -2.70$; $q_2 = -2.2$

$q_3 = 0.03$; $p_1 = 1.0$; $p_2 = 1.18$; rms = 1.314

FIGURE CAPTIONS

Fig.1 (a) The genetic code multiplicity can be obtained from the $Sp(6) \supset Sp(4) \otimes SU(2) \supset SU(2) \otimes SU(2) \otimes SU(2) \supset SU(2) \otimes \tilde{0}(2) \otimes SU(2) \supset SU(2) \otimes \tilde{0}(2) \otimes 0(2)$. The last $SU(2)$ breaking is partial. A tentative assignment of the primordial amino acids is suggested in set p I.

(b) The total break of the last $SU(2)$ in $0(2)$. In this case Phe, Ser, Arg and Cys would further subdivide leading to a code with 26 amino acids and the termination code.

Fig.2 (a) $Sp(4) \supset SU(2) \times SU(2) \supset \tilde{0}(2) \times SU(2) \supset S\tilde{0}(2) \otimes S\tilde{0}(2)$ chain, with a partial break at the last step to obtain a multiplicity "comparable" to the genetic code.

(b) $Sp(6) \supset SU(2) \otimes SU(2) \supset \tilde{0}(2) \times SU(2) \supset\supset \tilde{0}(2) \times \tilde{0}(2)$. Even though the chain ends with 20 amino acids there is no ordered way to break the symmetry and adjust the multiplicities to the ones of the universal genetic code.

Fig.3 Comparation between calculated and Grantham experimental polarities. Full bar are experimental and half fill bars the calculated values. The amino acids are numbered according to decreasing experimental polarities.

The numbers in the x-axis correspond to the amino acids as follows: 1 – Asp; 2 – Glu; 3 – Lys; 4 – Asn; 5 – Arg; 6 – Gln; 7 – His; 8 – Gly; 9 – Ser; 10 – Ala; 11 – Thr; 12 – Pro; 13 – Val; 14 – Tyr; 15 – Trp; 16 – Met; 17 – Phe; 18 – Leu; 19 – Ile; 20 – Cys.

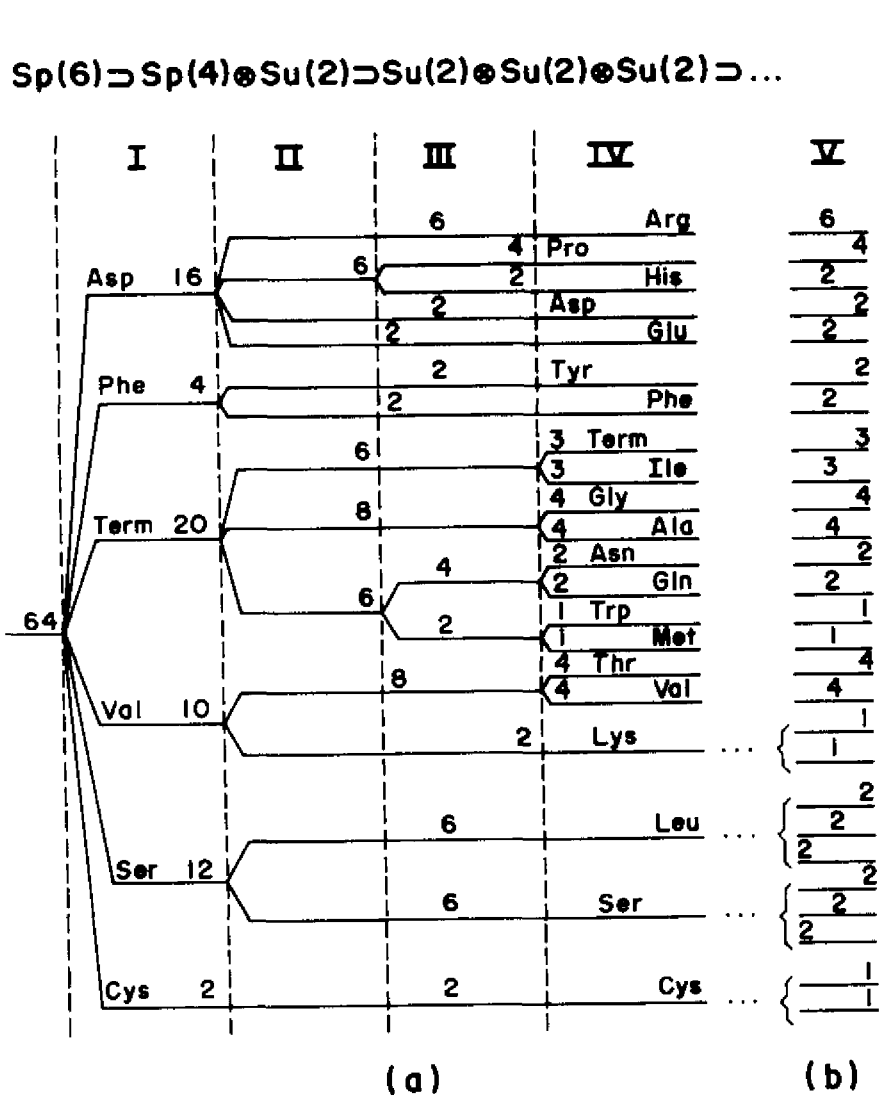

Sp(6)⊃Sp(4)⊛Su(2)⊃Su(2)⊛Su(2)⊛Su(2)⊃...
I
II
III
IV
V
64
Asp 16
Phe 4
Term 20
Val 10
Ser 12
Cys 2
Arg
Pro
His
Asp
Glu
Tyr
Phe
Term
Ile
Gly
Ala
Asn
Gln
Trp
Met
Thr
Val
Lys
Leu
Ser
Cys
(a)
(b)

Fig.1

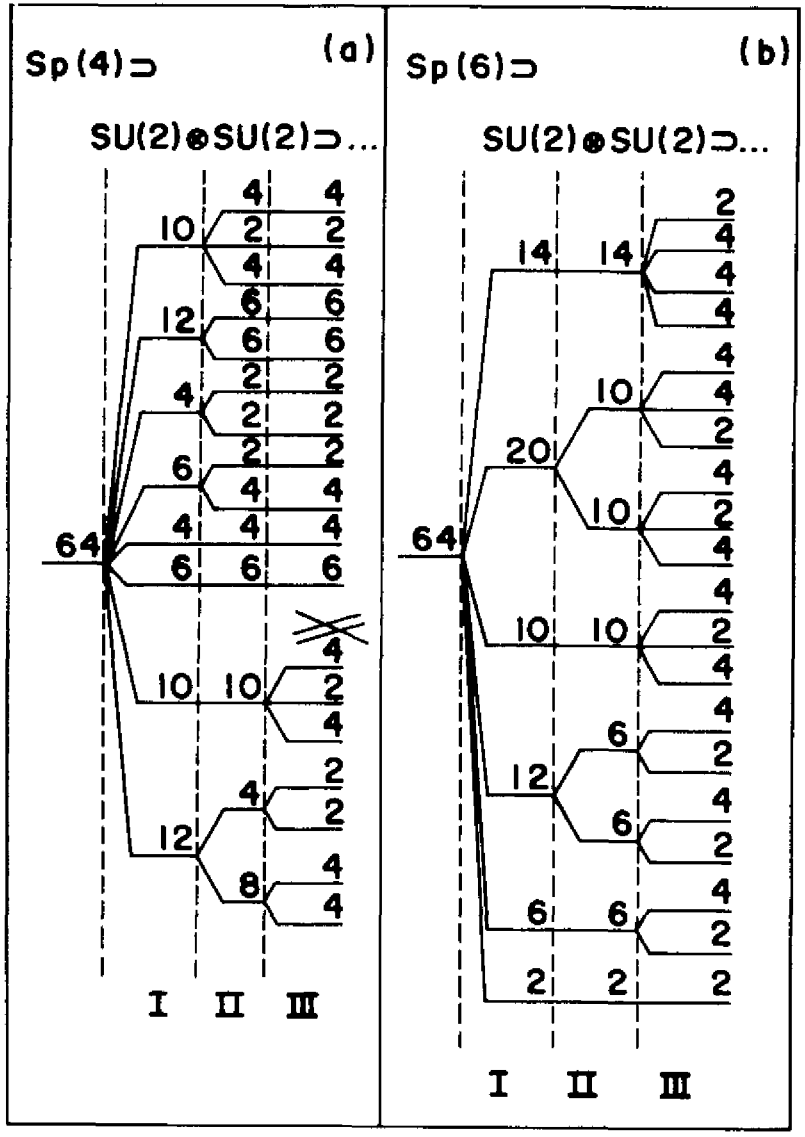

Sp(4)⊃
(a)
SU(2)⊛SU(2)⊃...
64
I
II
III
Sp(6)⊃
(b)
SU(2)⊛SU(2)⊃...
64
I
II
III

Fig.2

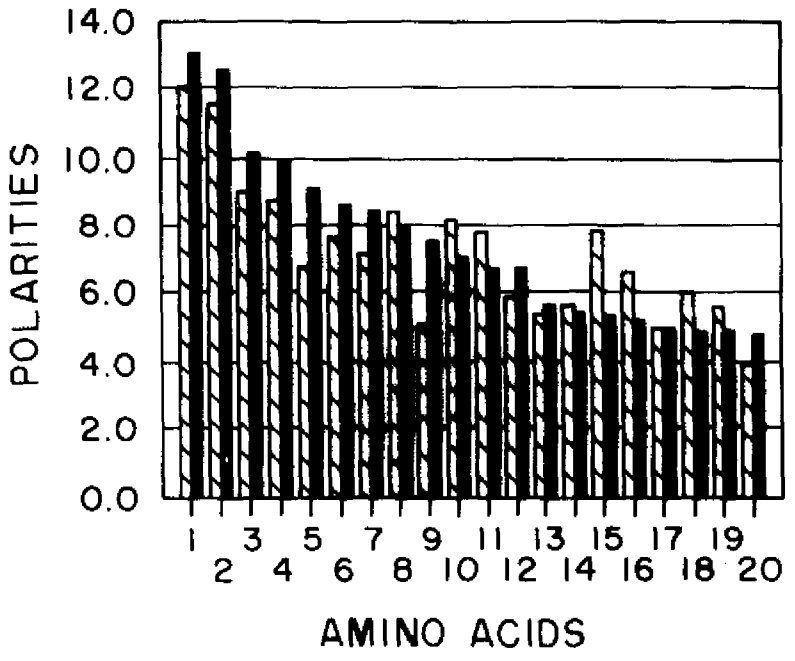

14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
POLARITIES
1 3 5 7 9 11 13 15 17 19
2 4 6 8 10 12 14 16 18 20
AMINO ACIDS

Fig.3

Stampato in proprio nella tipografia
del Centro Internazionale di Fisica Teorica