Proceedings of Institute of Mathematics of NAS of Ukraine 2000, Vol. 30, Part 2, 481–488.

Crystal Basis Model of the Genetic Code: Structure and Consequences

L. FRAPPAT †, *A. SCIARRINO* ‡ *and P. SORBA* †

† *Laboratoire d'Annecy-le-Vieux de Physique Théorique LAPTH CNRS, URA 1436, associée à l'Université de Savoie LAPP, BP 110, F-74941 Annecy-le-Vieux Cedex, France*
E-mail: frappat@lapp.in2p3.fr and sorba@lapp.in2p3.fr

‡ *Dipartimento di Scienze Fisiche, Università di Napoli "Federico II" and I.N.F.N., Sezione di Napoli, Mostra d'Oltremare, Pad. 20, I-80125 Napoli, Italy*
E-mail: sciarrino@na.infn.it

The main features of a model of genetic code based on the crystal basis of $U_{q\to 0}(sl(2)\oplus sl(2))$ is presented. The experimentally observed correlation between the values of the codon usage in quartets and sextets fits naturally in the model.

1 Introduction

Let us, briefly, remind how the DNA rules the synthesis of proteins, which constitute the most abundant organic substances in living matter systems. The DNA macromolecule is made of two linear chains of nucleotides wrapped in a double helix structure. Each nucleotide is characterized by one of the four elementary bases: adenine (A) and guanine (G) deriving from purine, and cytosine (C) and thymine (T) coming from pyrimidine. The DNA is localized in the nucleus of the cell and the transmission of the genetic information in the cytoplasm is achieved , schematically speaking, by the ribonucleic acid or RNA. This operation is called the transcription, the A, G, C, T bases in the DNA being respectively associated in RNA to the U, C, G, A bases, U denoting the uracile base. The correspondence law between triples of nucleotides, called *codons*, in the desoxyribonucleic acid (DNA) sequence and the amino-acids is called the genetic code. As a codon is an ordered sequence of three bases (e.g. AAG, AGA, etc.), obviously there are $4^3 = 64$ and different codons. Except the three following triples UAA, UAG and UGA, each of the 61 others is related through a ribosome to an amino-acid (a.a). In the universal eukariotic code, which constitutes the so called universal genetic code, the correspondence is given in Table 1. Thus the chain of nucleotides in the RNA – and also in the DNA – can also be viewed as a sequence of triples, each corresponding to aa a.a., except the three above mentioned codons. These last codons are called *Nonsense* or *Stop* codons, and their role is to stop the biosynthesis.

One can distinguish 20 different amino-acids: Alanine (Ala), Arginine (Arg), Asparagine (Asn), Aspartic acid (Asp), Cysteine (Cys), Glutamine (Gln), Glutamic acid (Glu), Glycine (Gly), Histidine (His), Isoleucine (Ile), Leucine (Leu), Lysine (Lys), Methionine (Met), Phenylalanine (Phe), Proline (Pro), Serine (Ser), Threonine (Thr), Tryptophane (Trp), Tyrosine (Tyr), Valine (Val). It follows that the different codons are associated to the same a.a., i.e. the genetic code is degenerated.

For the eukariotic code (see Table 1), the codons are organized in the following pattern of multiplets, each multiplet orresponding to a specific amino-acid:

1. 3 sextets: Arg, Leu, Ser
2. 5 quadruplets: Ala, Gly, Pro, Thr, Val
3. 2 triplets: Ile, Stop
4. 9 doublets: Asn, Asp, Cys, Gln, Glu, His, Lys, Phe, Tyr
5. 2 singlets: Met, Trp

It is natural, but not at all trivial, to ask if symmetry consideration can explain the existence of such an intriguing degenerate pattern. In our approach [1, 2] we consider the 4 nucleotides as elementary constituents of the codons. Actually, this approach mimicks the group theoretical classification of baryons made out from three quarks in elementary particles physics, the building blocks being here the A, C, G, T/U nucleotides. The main and essential difference stands in the property of a codon to be an *ordered* set of three nucleotides, which is not the case for a baryon. For example, there are three different codons made of the A, A, U nucleotides, namely AAU, AUA and UAA, while the proton appears as a weighted combination of the two u quarks and one d quark, that is $|p\rangle \sim |uud\rangle + |udu\rangle + |duu\rangle$. Constructing such pure states is made possible in the framework of the *crystal bases*, which can be defined in the limit $q \to 0$ of the deformation $\mathcal{U}_q(\mathcal{G})$ of any (semi)-simple classical Lie algebra $\mathcal{G}$.

2 The model

Introducing in $\mathcal{U}_{q\to 0}(\mathcal{G})$ the operators $\tilde{e}_i$ and $\tilde{f}_i$ ($i = 1, \ldots,$ rank $\mathcal{G}$), whose action on the elements of $\mathcal{U}_q(\mathcal{G})$-module is well-defined in the limit $q \to 0$, a particular kind of basis in a $\mathcal{U}_q(\mathcal{G})$-module can be defined [3]. Such a basis is called a crystal basis and carries the property to undergo in a specially simple way the action of the $\tilde{e}_i$ and $\tilde{f}_i$ operators: as an example, for any couple of vectors u, v in the crystal basis $\mathcal{B}$, one gets $u = \tilde{e}_i v$ if and only if $v = \tilde{f}_i u$. More interesting for our purpose is the property that, in the crystal basis, the basis vectors of the tensor product of two irreducible representations are *pure states*, [3]. Let us emphasize once more the motivation for our choice of the crystal basis. It is an observed fact that in the codons the order of the nucleotides is of fundamental importance (e.g. CCU $\to$ `Pro`, CUC $\to$ `Leu`, UCC $\to$ `Ser`). We want to consider the codons as composite states of the (elementary) nucleotides, but this surely cannot be done in the framework of Lie (super)algebras. Indeed in the Lie theory the composite states, obtained by the tensor product of the fundamental irreducible rpresentations, are linear combinations of the elementary states, with symmetry properties determined by the tensor product (i.e. for $sl(n)$ by the structure of the corresponding Young tableau). The crystal basis on the contrary provides us with the mathematical structure to build composite states as *pure* states, characterized by the order of the constituents. In order to dispose of such a basis, we need to consider the limit $q \to 0$. Note that in this limit we do not deal anymore with a Lie algebra either with an universal deformed enveloping algebra.

We consider the four nucleotides as basic states of the $(\frac{1}{2}, \frac{1}{2})$ representation of the $\mathcal{U}_q(sl(2) \oplus sl(2))$ quantum enveloping algebra in the limit $q \to 0$. A triplet of nucleotides will then be obtained by constructing the tensor product of three such four dimensional representations. The algebra $\mathcal{G} = su(2) \oplus su(2))$ appears the most natural for our purpose. First of all, it is "reasonable" to represent the four nucleotides in the fundamental representation of $\mathcal{G}$. Moreover, the complementary rule in the DNA–RNA transcription may suggest to assign a *quantum number* with opposite values to the couples (A,T/U) and (C,G). The distinction between the purine bases (A,G) and the pyrimidine ones (C,T/U) can be algebraically represented in an analogous way. Thus considering the representation $\left(\frac{1}{2}, \frac{1}{2}\right)$ of the group $SU(2) \times SU(2)$ and denoting $\pm$ the

basis vector corresponding to the eigenvalues $\pm\frac{1}{2}$ of the J_3 generator in any of the two $su(2)$ corresponding algebras, we will assume the following "biological" spin structure:

$$\begin{array}{ccc} & \overset{su(2)_H}{} & \\ C \equiv (+,+) & \longleftrightarrow & U \equiv (-,+) \\ su(2)_V \updownarrow & & \updownarrow su(2)_V \\ G \equiv (+,-) & \longleftrightarrow & A \equiv (-,-) \\ & \underset{su(2)_H}{} & \end{array} \tag{1}$$

the subscripts H (:= horizontal) and V (:= vertical) being just added to specify the group actions.

To represent a codon, we will have to perform the tensor product of three $\left(\frac{1}{2},\frac{1}{2}\right)$ representations of $\mathcal{U}_{q\to 0}(sl(2)\oplus sl(2))$. We get, using the Kashiwara theorem [3], the following tables

$$\left(\tfrac{3}{2},\tfrac{3}{2}\right) \equiv \begin{pmatrix} \text{CCC} & \text{UCC} & \text{UUC} & \text{UUU} \\ \text{GCC} & \text{ACC} & \text{AUC} & \text{AUU} \\ \text{GGC} & \text{AGC} & \text{AAC} & \text{AAU} \\ \text{GGG} & \text{AGG} & \text{AAG} & \text{AAA} \end{pmatrix}$$

$$\left(\tfrac{3}{2},\tfrac{1}{2}\right)^1 \equiv \begin{pmatrix} \text{CCG} & \text{UCG} & \text{UUG} & \text{UUA} \\ \text{GCG} & \text{ACG} & \text{AUG} & \text{AUA} \end{pmatrix}$$

$$\left(\tfrac{3}{2},\tfrac{1}{2}\right)^2 \equiv \begin{pmatrix} \text{CGC} & \text{UGC} & \text{UAC} & \text{UAU} \\ \text{CGG} & \text{UGG} & \text{UAG} & \text{UAA} \end{pmatrix}$$

$$\left(\tfrac{1}{2},\tfrac{3}{2}\right)^1 \equiv \begin{pmatrix} \text{CCU} & \text{UCU} \\ \text{GCU} & \text{ACU} \\ \text{GGU} & \text{AGU} \\ \text{GGA} & \text{AGA} \end{pmatrix} \qquad \left(\tfrac{1}{2},\tfrac{3}{2}\right)^2 \equiv \begin{pmatrix} \text{CUC} & \text{CUU} \\ \text{GUC} & \text{GUU} \\ \text{GAC} & \text{GAU} \\ \text{GAG} & \text{GAA} \end{pmatrix}$$

$$\left(\tfrac{1}{2},\tfrac{1}{2}\right)^1 \equiv \begin{pmatrix} \text{CCA} & \text{UCA} \\ \text{GCA} & \text{ACA} \end{pmatrix} \qquad \left(\tfrac{1}{2},\tfrac{1}{2}\right)^2 \equiv \begin{pmatrix} \text{CGU} & \text{UGU} \\ \text{CGA} & \text{UGA} \end{pmatrix}$$

$$\left(\tfrac{1}{2},\tfrac{1}{2}\right)^3 \equiv \begin{pmatrix} \text{CUG} & \text{CUA} \\ \text{GUG} & \text{GUA} \end{pmatrix} \qquad \left(\tfrac{1}{2},\tfrac{1}{2}\right)^4 \equiv \begin{pmatrix} \text{CAC} & \text{CAU} \\ \text{CAG} & \text{CAA} \end{pmatrix}$$

3 The Reading (or Ribosome) operator $\mathcal{R}$

Our model does not gather codons associated to one particular a.a. in the same irreducible multiplet. However, it is possible to construct an operator $\mathcal{R}$ out of the algebra $\mathcal{U}_{q\to 0}(sl(2)\oplus sl(2))$, acting on the codons, that will describe the Evarious genetic codes in the following way:

Two codons have the same eigenvalue under $\mathcal{R}$ if and only if they are associated to the same amino-acid. This operator $\mathcal{R}$ will be called the reading operator. It is possible to construct a $\mathcal{R}$ for the various genetic codes. Here we limit orselves to present in detail only the Reading operator for the Eukaryotic code

$$\begin{aligned} \mathcal{R}_{EC} = {} & \tfrac{4}{3}c_1\, C_H + \tfrac{4}{3}c_2\, C_V - 4c_1\, \mathcal{P}_H\, J_{H,3} - 4c_2\, \mathcal{P}_V\, J_{V,3} + (-8c_1\, \mathcal{P}_D + (8c_1+12c_2)\, \mathcal{P}_S)\, J_{V,3} \\ & + (-4c_1+14c_2)\, \mathcal{P}_{AG}\left(\tfrac{1}{2} - J^{(3)}_{V,3}\right) \\ & + \left[12c_2\, \mathcal{P}_{AU} + (6c_1+6c_2)\, \mathcal{P}_{UG}\right]\left(\tfrac{1}{2} - J^{(3)}_{V,3}\right)\left(\tfrac{1}{2} - J^{(3)}_{H,3}\right), \end{aligned} \tag{2}$$

where

- the operators $J_{\alpha,3}$ ($\alpha = H, V$) are the third components of the total spin generators of the algebra $\mathcal{U}_{q\to 0}(sl(2) \oplus sl(2))$;
- the operator C_α is a Casimir operator of $\mathcal{U}_{q\to 0}(sl(2))$ in the crystal basis. It commutes with $J_{\alpha\pm}$ and $J_{\alpha,3}$ (where $J_{\alpha\pm}$ are the generators with a well-defined behaviour for $q \to 0$) and its eigenvalues on any vector basis of an irreducible representation of highest weight J is $J(J+1)$, that is the same as the undeformed standard second degree Casimir operator of $sl(2)$. Its explicit expression is

$$C_\alpha = (J_{\alpha,3})^2 + \tfrac{1}{2} \sum_{n\in\mathbb{Z}_+} \sum_{k=0}^{n} (J_{\alpha-})^{n-k} (J_{\alpha+})^n (J_{\alpha-})^k; \tag{3}$$

- $\mathcal{P}_H$, $\mathcal{P}_V$, $\mathcal{P}_D$, $\mathcal{P}_S$, $\mathcal{P}_{AG}$, $\mathcal{P}_{AU}$ and $\mathcal{P}_{UG}$ are projectors operators given by:

$$\mathcal{P}_H = J^d_{H+}\, J^d_{H-} \qquad \text{and} \qquad \mathcal{P}_V = J^d_{V+}\, J^d_{V-}, \tag{4}$$

$$\begin{aligned} \mathcal{P}_D = {} & \left(1 - J^d_{V+}\, J^d_{V-}\right) \left(J^d_{H+}\, J^d_{H-}\right) \left(J^d_{H-}\, J^d_{H+}\right) \\ & + \left(1 - J^d_{H+}\, J^d_{H-}\right) \left(1 - J^d_{V+}\, J^d_{V-} \left(1 - J^d_{H-}\, J^d_{H+}\right)\right), \end{aligned} \tag{5}$$

$$\begin{aligned} \mathcal{P}_S = {} & \left(J^d_{H-}\, J^d_{H+}\right) \left[\left(J^d_{H+}\, J^d_{H-}\right) \left(1 - J^d_{V+}\, J^d_{V-}\right)\right. \\ & \left. + \left(J^d_{V+}\, J^d_{V-}\right) \left(J^d_{V-}\, J^d_{V+}\right) \left(1 - J^d_{H+}\, J^d_{H-}\right)\right], \end{aligned} \tag{6}$$

$$\mathcal{P}_{AG} = \left(J^d_{H+}\, J^d_{H-})(J^d_{H-}\, J^d_{H+}\right) \left(1 - J^d_{V+}\, J^d_{V-}\right) \left(J^d_{V-}\, J^d_{V+}\right), \tag{7}$$

$$\mathcal{P}_{AU} = \left(1 - J^d_{H+}\, J^d_{H-}\right) \left(J^d_{H-}\, J^d_{H+}\right) \left(J^d_{V+}\, J^d_{V-}\right) \left(J^d_{V-}\, J^d_{V+}\right), \tag{8}$$

$$\mathcal{P}_{UG} = \left(J^d_{H+}\, J^d_{H-}\right) \left(J^d_{H-}\, J^d_{H+}\right) \left(1 - J^d_{V+}\, J^d_{V-}\right) \left(1 - J^d_{V-}\, J^d_{V+}\right). \tag{9}$$

We get the following eigenvalues of the reading operators for the amino-acids (after a rescaling, setting $c \equiv c_1/c_2$):

a.a.	value of $\mathcal{R}$	a.a.	value of $\mathcal{R}$	a.a.	value of $\mathcal{R}$
Ala	$-c+3$	Gly	$-c+5$	Pro	$-c-1$
Arg	$-c+1$	His	$-3c+1$	Ser	$3c-1$
Asn	$9c+5$	Ile	$5c+9$	Thr	$3c+3$
Asp	$5c+5$	Leu	$c-1$	Trp	$3c-5$
Cys	$3c+7$	Lys	$17c+5$	Tyr	$c+1$
Gln	$5c+1$	Met	$5c-3$	Val	$c+3$
Glu	$13c+5$	Phe	$-7c-1$	Ter	$9c+1$

(10)

Remark that the reading operators $\mathcal{R}(c)$ can be used for any real value of c, except a finite set of rational values confering the same eigenvalue to codons relative to two different amino-acids. Moreover from our algebra it is possible to construct a hamiltonian which gives a very satisfactory fit of the 16 values of the free energy released in the folding of RNA [1].

codon	a.a.	J_H	J_V	codon	a.a.	J_H	J_V
CCC	Pro	3/2	3/2	UCC	Ser	3/2	3/2
CCU	Pro	$(1/2$	$3/2)^1$	UCU	Ser	$(1/2$	$3/2)^1$
CCG	Pro	$(3/2$	$1/2)^1$	UCG	Ser	$(3/2$	$1/2)^1$
CCA	Pro	$(1/2$	$1/2)^1$	UCA	Ser	$(1/2$	$1/2)^1$
CUC	Leu	$(1/2$	$3/2)^2$	UUC	Phe	3/2	3/2
CUU	Leu	$(1/2$	$3/2)^2$	UUU	Phe	3/2	3/2
CUG	Leu	$(1/2$	$1/2)^3$	UUG	Leu	$(3/2$	$1/2)^1$
CUA	Leu	$(1/2$	$1/2)^3$	UUA	Leu	$(3/2$	$1/2)^1$
CGC	Arg	$(3/2$	$1/2)^2$	UGC	Cys	$(3/2$	$1/2)^2$
CGU	Arg	$(1/2$	$1/2)^2$	UGU	Cys	$(1/2$	$1/2)^2$
CGG	Arg	$(3/2$	$1/2)^2$	UGG	Trp	$(3/2$	$1/2)^2$
CGA	Arg	$(1/2$	$1/2)^2$	UGA	Ter	$(1/2$	$1/2)^2$
CAC	His	$(1/2$	$1/2)^4$	UAC	Tyr	$(3/2$	$1/2)^2$
CAU	His	$(1/2$	$1/2)^4$	UAU	Tyr	$(3/2$	$1/2)^2$
CAG	Gln	$(1/2$	$1/2)^4$	UAG	Ter	$(3/2$	$1/2)^2$
CAA	Gln	$(1/2$	$1/2)^4$	UAA	Ter	$(3/2$	$1/2)^2$
GCC	Ala	3/2	3/2	ACC	Thr	3/2	3/2
GCU	Ala	$(1/2$	$3/2)^1$	ACU	Thr	$(1/2$	$3/2)^1$
GCG	Ala	$(3/2$	$1/2)^1$	ACG	Thr	$(3/2$	$1/2)^1$
GCA	Ala	$(1/2$	$1/2)^1$	ACA	Thr	$(1/2$	$1/2)^1$
GUC	Val	$(1/2$	$3/2)^2$	AUC	Ile	3/2	3/2
GUU	Val	$(1/2$	$3/2)^2$	AUU	Ile	3/2	3/2
GUG	Val	$(1/2$	$1/2)^3$	AUG	Met	$(3/2$	$1/2)^1$
GUA	Val	$(1/2$	$1/2)^3$	AUA	Ile	$(3/2$	$1/2)^1$
GGC	Gly	3/2	3/2	AGC	Ser	3/2	3/2
GGU	Gly	$(1/2$	$3/2)^1$	AGU	Ser	$(1/2$	$3/2)^1$
GGG	Gly	3/2	3/2	AGG	Arg	3/2	3/2
GGA	Gly	$(1/2$	$3/2)^1$	AGA	Arg	$(1/2$	$3/2)^1$
GAC	Asp	$(1/2$	$3/2)^2$	AAC	Asn	3/2	3/2
GAU	Asp	$(1/2$	$3/2)^2$	AAU	Asn	3/2	3/2
GAG	Glu	$(1/2$	$3/2)^2$	AAG	Lys	3/2	3/2
GAA	Glu	$(1/2$	$3/2)^2$	AAA	Lys	3/2	3/2

Table 1: The eukariotic code. The upper label denotes different IR.

	Biological organism	Type	number of sequences	number of codons
1	Homo sapiens	v	14 529	7 168 914
2	Saccharomyces cerevisiae	f	11 771	5 691 597
3	Caenorhabditis elegans	i	12 638	5 514 021
4	Rattus norvegicus	v	4 430	2 135 734
5	Arabidopsis Thaliana	p	3 533	1 497 366
6	Drosophila melanogaster	i	2 625	1 443 176
7	Schizosaccharomyces pombe	f	2 289	1 093 794
8	Gallus gallus	v	1 454	701 782
9	Xenopus laevis	v	1 255	551 494
10	Bos taurus	v	1 217	528 790
11	Oryctolagus cuniculus	v	674	335 049
12	Sus scrofa	v	589	238 579
13	Zea mays	p	603	222 493

Table 2: v) Vertebrates – i) Invertebrata – p) Plants – f) Fungi

Species	Pro	Ala	Thr	Ser	Val	Leu	Arg	Gly
Homo sapiens	2,36	2,05	2,26	2,52	0,23	0,16	0,53	1,00
Saccharomyces c.	3,44	2,64	2,22	2,19	1,10	1,28	1,73	1,83
Caenorhabditis e.	3,17	2,78	2,48	1,88	0,74	0,69	2,82	7,80
Rattus Norveg.	2,45	2,18	2,34	2,34	0,22	0,17	0,62	1,04
Arabidopsis Thal.	1,93	1,97	2,03	1,95	0,53	0,96	1,29	2,45
Drosophila mel.	0,78	0,89	0,75	0,41	0,21	0,18	0,96	4,02
Schizosaccharomyces	2,74	2,94	2,12	2,25	1,49	1,39	2,63	3,66
Gallus gallus	1,82	1,92	1,97	1,90	0,25	0,14	0,51	1,02
Xenopus laevis	4,09	4,32	4,04	3,39	0,48	0,32	1,00	1,70
Bos taurus	1,88	1,62	1,77	2,03	0,19	0,13	0,51	0,95
Oryctolagus cun.	1,51	1,49	1,31	1,50	0,15	0,10	0,45	0,88
Sus scrofa	1,59	1,59	1,50	1,62	0,16	0,12	0,45	0,89
Zea mays	0,87	0,69	0,83	0,98	0,19	0,24	0,39	0,85

Table 3: Branching ratio B_{AG}

Species	Pro	Ala	Thr	Ser	Val	Leu	Arg	Gly
Homo sapiens	2,45	2,44	1,96	3,22	0,36	0,30	0,41	0,66
Saccharomyces c.	2,57	3,44	2,54	2,75	2,06	1,17	3,73	4,00
Caenorhabditis e.	1,07	3,14	2,38	1,58	1,85	1,97	2,78	2,67
Rattus Norveg.	2,67	2,84	1,99	3,28	0,32	0,28	0,48	0,71
Arabidopsis Thal.	2,21	3,34	2,46	2,78	1,53	2,48	2,06	2,30
Drosophila mel.	0,40	1,03	0,63	0,37	0,38	0,22	1,12	3,13
Schizosaccharomyces	4,75	5,70	3,52	3,85	3,58	4,08	5,48	5,18
Gallus gallus	1,72	2,24	1,62	2,32	0,42	0,27	0,57	0,66
Xenopus laevis	3,63	4,68	3,57	4,80	0,73	0,59	1,12	1,06
Bos taurus	1,96	2,10	1,52	2,72	0,32	0,26	0,38	0,65
Oryctolagus cun.	1,63	1,82	1,18	2,03	0,28	0,21	0,34	0,53
Sus scrofa	1,72	2,10	1,43	2,42	0,27	0,23	0,36	0,59
Zea mays	0,78	1,00	0,94	1,08	0,54	0,59	0,67	1,06

Table 4: Branching ratio B_{UG}

Species	Pro	Ala	Thr	Ser	Val	Leu	Arg	Gly
Homo sapiens	2,90	3,82	3,13	3,99	0,51	0,49	0,96	1,41
Saccharomyces c.	1,29	2,06	1,58	1,66	1,09	0,51	1,49	1,62
Caenorhabditis e.	0,49	1,65	1,25	0,93	0,98	1,29	1,22	1,43
Rattus Norveg.	2,93	4,12	3,25	4,20	0,54	0,50	1,00	1,47
Arabidopsis Thal.	0,66	1,24	1,45	1,25	0,74	1,65	0,80	0,90
Drosophila mel.	1,12	2,52	1,59	1,17	0,54	0,36	2,38	6,53
Schizosaccharomyces	1,80	2,23	1,67	1,51	1,35	1,18	2,08	1,99
Gallus gallus	2,29	2,73	2,33	3,03	0,50	0,43	1,25	1,29
Xenopus laevis	2,85	4,02	3,38	4,09	0,58	0,48	1,14	1,17
Bos taurus	2,70	3,76	2,84	3,76	0,53	0,48	1,00	1,46
Oryctolagus cun.	2,58	3,83	2,51	3,70	0,54	0,48	1,18	1,55
Sus scrofa	2,58	3,94	2,95	3,75	0,56	0,50	1,05	1,56
Zea mays	0,90	1,48	1,79	1,70	0,80	0,98	1,75	2,22

Table 5: Branching ratio B_{CG}

4 Correlations of codon usage

In the following the labels X, Y, Z, V represent any of the 4 bases C, U, G, A. Let XYZ be a codon in a given multiplet, say m_i, encoding an a.a., say A_i. We define the probability of usage of the codon XYZ as the ratio between the frequency of usage n_Z of the codon XYZ in the biosynthesis of A_i and the total number N of synthetized A_i, i.e. as the relative codon frequency, in the limit of *very large* N. The frequency rate of usage of a codon in a multiplet is connected to its probability of usage $P(XYZ \to \text{a.a.})$. We define the *branching ratio* B_{ZV} as

$$B_{ZV} = \frac{P(XYZ \to A_i)}{P(XYV \to A_i)}, \tag{11}$$

where XYV is another codon belonging to the same multiplet m_i. It sounds reasonable to argue that in the limit of very large number of codons, for a fixed biological organism and amino-acid, the branching ratio depends essentially on the properties of the codon. In our model this means that in this limit B_{ZV} is a function, depending on the type of the multiplet, on the *quantum numbers* of the codons XYZ and XYV, i.e. on the labels J_α, J^3_α, and on an other set of quantum labels leaving out the degeneracy on J_α; in Table 1 different irreducible representations with the same values of J_α are distinguished by an upper label. Moreover we assume that B_{ZV}, in the limit above specified, depends only on the irreducible representation (IR) of the codons, i.e.:

$$B_{ZV} = F_{ZV}(b.o.; IR(XYZ); IR(XYV)), \tag{12}$$

where we have explicitly denoted by $b.o.$ the dependence on the biological species. Let us point out that the branching ratio has a meaning only if the codons XYZ and XYU are in the same multiplet, i.e. if they code the same amino-acid.

In the following, we consider the quartets and the quartet sub-parts of the sextets, i.e. the 4 codons which differ only for the codon in third position. There are five quartets and three sextets in the eukariotic code: that will allow a rather detailed analysis. We recall that the 5 amino-acids coded by the quartets are `Pro`, `Ala`, `Thr`, `Gly` ,`Val` and the 3 amino-acids coded by the sextets are `Leu`, `Arg`, `Ser`. There are, for the quartets, 6 branching ratios, of which only 3 are independent. We choose as fundamental ones the ratios B_{AG}, B_{CG} and B_{UG}. It happens that we can define several functions B_{ZV}, considering ratios of probability of codons differing for the first two nucleotides XY, i.e.

$$\begin{aligned} B_{ZV} &= F_{ZV}(b.o.; IR(XYZ); IR(XYV)), \\ B'_{ZV} &= F_{ZV}(b.o.; IR(X'Y'Z); IR(X'Y'V)). \end{aligned} \tag{13}$$

Then if the codon XYZ (XYV) and $X'Y'Z$ ($X'Y'V$) are respectively in the same irreducible representation, it follows that

$$B_{ZV} = B'_{ZV}. \tag{14}$$

The analysis was performed on a set of data retrieved (May 1999) from the data bank of "Codon usage tabulated from GenBank" [4]. In particular in [5] we analyzed the data set with more than 64.000 codons and we found 34 biological species (neglecting 3 biological species belonging to protozoo, bacteries and mushrooms). This has to be compared with the result of [2] where such a correlation has been remarked for 12 biological species belonging only to the vertebrate series. Here we present th results only for the subset of 13 species with more than 200.000 codons, see Table 2.

In Table 3, 4 and 5 the B_{AG}, B_{UG} and B_{CG} are reported for the 13 amino-acids coded by the quartets and sextets showing:

- a clear correlation between the four amino-acids `Pro`, `Ala`, `Thr` and `Ser`. From Table 1 we see that for these amino-acids the irreducible representation involved in the numerator of the branching ratios (see (11)) is always the same: $(1/2,1/2)^1$ for B_{AG}, $(1/2,3/2)^1$ for B_{UG}, $(3/2,3/2)$ for B_{CG}, while the irreducible representation in the denominator is $(3/2,1/2)^1$ for the whole set.
- a clear correlation between the two amino-acids `Val` and `Leu`. From Table 1 we see that also for these two amino-acids the irreducible representation in the numerator of (11) is the same: $(1/2,1/2)^3$ for B_{AG}, $(1/2,3/2)^2$ for B_{UG}, $(1/2,3/2)^2$ for B_{CG}, and the irreducible representation in the denominator is $(1/2,1/2)^3$.
- no correlation of the `Arg` and also of the `Gly` with the others amino-acids, in agreement with the irreducible representation assignment of Table 1.

5 Conclusion

The model we propose is based on symmetry principles. The symmetry algebra $\mathcal{U}_{q\to 0}(sl(2)\oplus sl(2))$ that we have chosen has two main characteristics. First it encodes the stereochemical property of a base, and also reflects the complementarity rule, by confering quantum numbers to each nucleotide. Secondly, it admits representation spaces or crystal bases in which an ordered sequence of nucleotides or codon can be suitably characterized. Let us add that it is a remarkable property of a quantum algebra in the limit $q \to 0$ to admit representations, obtained from the tensorial product of basic ones, in which each state appears as a unique sequence of ordered basic elements. In this framework, the correspondence codon/amino-acid is realized by the operator $\mathcal{R}_c$, constructed out of the symmetry algebra, and acting on codons: the eigenvalues provided by $\mathcal{R}_c$ on two codons will be the same or different following the two codons are associated to the same or to two different amino-acids. The model does not necessarily assign the codons in a multiplet (in particular the quartets, sextets and triplet) to the same irreducible representation. This feature is relevant as it may explain the different codon usage between codons encoding the same a.a.. Indeed, as we have shown in this paper, it fits very well with our model the observed fact that for *any biological organism*, in the limit of large number of biosynthetized amino-acids, the ratios B_{AG}, B_{UG} and B_{CG} for, Pro, Ala, Thr, Ser, in one side, and Val, Leu, in other side, are very close. Let us remark that obviously these ratios depend on the biological organism and we are unable to make any prevision on their values, but only that their values should be correlated.

References

[1] Frappat L., Sciarrino A. and Sorba P., A crystal base for the genetic code, *Phys. Lett. A*, 1998, V.250, 214; `physics/9801027`.

[2] Frappat L., Sciarrino A. and Sorba P., Symmetry and codon usage correlations in the genetic code, preprint LAPTH-709/98; `physics/9812041`; *Phys. Lett. A.*, in publication.

[3] Kashiwara M., Crystallizing the q-analogue of universal enveloping algebras, *Commun. Math. Phys.*, 1990, V.133, 249.

[4] Nakamura Y., Gojobori T. and Ikemura T., Nucleic Acids Research, 1998, V.26, 334.

[5] Chiusano M.L., Frappat L., Sciarrino A. and Sorba P., Codon usage correlations and Crystal Basis model of the genetic code, preprint LAPTH-736/99, DSF-Th-17/99.