

# Symmetry and Codon Usage Correlations in the Genetic Code

L. Frappat, P. Sorba

*Laboratoire de Physique Théorique LAPTH, URA 1436,  
Chemin de Bellevue, BP 110,  
F-74941 Annecy-le-Vieux, France  
E-mail: frappat(sorba)@lapp.in2p3.fr*

A. Sciarrino

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

## Abstract

The ratios of the codon usage in the quartets and sextets for the vertebrate series exhibit a correlated behaviour which fits naturally in the framework of the crystal basis model of the genetic code [1]. Moreover the observed universal behaviour of these suitably normalized ratios can be easily explained.



# 1 Introduction

It is a well known and intriguing fact that, in the genetic code, 64 codons code the biosynthesis of 20 amino-acids (a.a.) with a structure in multiplets reported in Table 1.

It is also a well known and, at our knowledge, unexplained fact that the frequency rate of usage (codon usage) of the different codons inside a multiplet is not the same.

It is the aim of this paper to emphasize for the vertebrate series a correlation in the codon usage, in the quartets and sextets, which is naturally explained in the framework of the mathematical model of the genetic code recently proposed by the authors [1]. Moreover we put in evidence an universal function behaviour connected with the codon usage, which also finds a justification in the model.

In Sec. 2 we recall the essential features of the model and in Sec. 3 we present the analysis of the codon usage for a set of biological sequences in the vertebrate series [3].

## 2 The crystal basis model

### 2.1 The crystal basis

Let us briefly recall some properties of the crystal basis [2]. We limit ourselves to the case of  $\mathcal{U}_q(sl(2))$ , but such basis exists for any finite dimensional representation of  $\mathcal{U}_{q \rightarrow 0}(\mathcal{G})$ ,  $\mathcal{G}$  being any (semi)-simple classical Lie algebra. The crystal basis has the nice property, for  $\mathcal{U}_q(sl(2))$ , that in the limit  $q \rightarrow 0$

$$\tilde{J}_+ u_k = u_{k+1} \quad \text{for } 0 \leq k < 2J \quad (1)$$

$$\tilde{J}_- u_k = u_{k-1} \quad \text{for } 0 < k \leq 2J \quad (2)$$

$$\tilde{J}_3 u_k = (k - J) u_k \quad \text{for } 0 \leq k \leq 2J \quad (3)$$

and

$$\tilde{J}_+ u_{2J} = \tilde{J}_- u_0 = 0 \quad (4)$$

where the operators  $\tilde{J}_\pm$  are a redefinition, using an element of the center, of the generators  $J_\pm$  of  $\mathcal{U}_q(sl(2))$ ,  $\tilde{J}_3 = J_3$ , and  $u_k$  are the basis vectors of the irreducible representation labelled by  $J$  ( $J$  being an integer or half-integer). The labels of the irreducible representation are connected to the eigenvalues of the ‘‘Casimir’’ operator  $C$ :

$$C = (\tilde{J}_3)^2 + \frac{1}{2} \sum_{n \in \mathbb{Z}_+} \sum_{k=0}^n (\tilde{J}_-)^{n-k} (\tilde{J}_+)^n (\tilde{J}_-)^k. \quad (5)$$

Its eigenvalue on any vector basis of the irreducible  $J$ -representation is  $J(J+1)$ .

Moreover any state in the tensor product of two irreducible representations  $R_1 \otimes R_2$  is written in the crystal basis as one and only one tensor product of a  $R_1$  state by a  $R_2$  state [2]. For example, taking for  $R_1$  and  $R_2$  the two-dimensional representation  $J = \frac{1}{2}$  of  $\mathcal{U}_{q \rightarrow 0}(sl(2))$  with states  $|+\rangle$  and  $|-\rangle$ , one will get in  $R_1 \otimes R_2$  the  $J = 1$  representation displayed by  $|+, +\rangle$ ,  $|-, +\rangle$  and  $|-, -\rangle$ , and the  $J = 0$  representation with the state  $|+, -\rangle$ .

Now let us state the main hypothesis of our model [1].

## 2.2 The assumptions

**Assumption I** - The four nucleotides containing the bases: adenine ( $A$ ) and guanine ( $G$ ), deriving from purine, and cytosine ( $C$ ) and thymine ( $T$ ), coming from pyrimidine, are the basis vectors of a crystal basis of the  $(1/2, 1/2)$  irreducible representation of the quantum algebra  $\mathcal{U}_q(sl(2) \oplus sl(2))$  in the limit  $q \rightarrow 0$ . In the following, we denote with  $\pm$  the basis vector corresponding to the eigenvalue  $\pm 1/2$  of  $J_\alpha^3$ , where  $\alpha = H$  ( $V$ ) specifies the generator of the first (second)  $sl(2)$ . We assume the following “spin” structure (we remind that the thymine  $T$  in the DNA is replaced by the uracile  $U$  in the RNA):

$$\begin{array}{ccc}
 & sl(2)_H & \\
 C \equiv (+, +) & \longleftrightarrow & U \equiv (-, +) \\
 & & \\
 & sl(2)_V \updownarrow & \\
 G \equiv (+, -) & \longleftrightarrow & A \equiv (-, -) \\
 & sl(2)_H &
 \end{array} \tag{6}$$

Let us remark that the  $H$ -symmetry is associated to the purine-pyrimidine structure, while the  $V$ -symmetry reflects the complementarity rule (that is  $A - T/U$  and  $C - G$  interactions).

**Assumption II** - The *codons* are the basis vectors, in the crystal basis, of the irreducible representations build up by the tensor product of three 4-dimensional  $(\frac{1}{2}, \frac{1}{2})$  fundamental representations describing the nucleotides.

We have reported in Table 1 the assignment of the codons classified in the representations which appear in the r.h.s. of the following relation:

$$\left(\frac{1}{2}, \frac{1}{2}\right) \otimes \left(\frac{1}{2}, \frac{1}{2}\right) \otimes \left(\frac{1}{2}, \frac{1}{2}\right) = \left(\frac{3}{2}, \frac{3}{2}\right) \oplus 2\left(\frac{3}{2}, \frac{1}{2}\right) \oplus 2\left(\frac{1}{2}, \frac{3}{2}\right) \oplus 4\left(\frac{1}{2}, \frac{1}{2}\right) \tag{7}$$

In [1], an operator (called the *reading or ribosome operator*)  $\mathcal{R}$  has been constructed out of the algebra  $\mathcal{U}_{q \rightarrow 0}(sl(2) \oplus sl(2))$ , which describes the multiplet structure of the the genetic code 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*. Moreover an “Hamiltonian” depending on 4 parameters has been build up which gives a very satisfactory fit of the 16 values of the free energy released in the folding of a RNA sequence into a base paired double helix.

Let us close this section by drawing the reader’s attention to Fig. 1 where is specified for each codon its position in the appropriate representation. The diagram of states for each representation is supposed to lie in a separate parallel plane. Thick lines connect codons associated to the same amino-acid. One remarks that each segment relates a couple of codons belonging to the same representation or to two different representations. This last case occurs for quadruplets or sextets of codons associated to the same amino-acid. It is the purpose of this letter to show a remarkable relation between such multiplets of codons (or amino-acids) involving the same subset of representation and (branching ratios of) the probabilities of presence of codons in the amino-acid biosynthesis.

### 3 Correlation of codon usage

We define the codon usage as the frequency of use of a given codon in the process of biosynthesis of all the amino-acids. We define the probability of usage of the codon  $XYZ$  of a given amino-acid as the ratio between the occurrence of the codon  $XYZ$  and the occurrence  $N$  of the corresponding amino-acid, i.e. as the relative codon frequency, in the limit of very large  $N$ . Here and in the following the labels  $X, Y, Z, V$  represent the bases  $C, U, G, A$ . The frequency rate of usage of a codon in a multiplet is connected to its probability of usage  $P(XYZ \rightarrow \text{a.a.})$ . It is reasonable to assume that  $P(XYZ \rightarrow \text{a.a.})$  depends on:

- the sequence analyzed
- the nature of the neighboring codons in the sequence
- the amino-acid (a.a.)
- the nature and structure of the multiplet associated to the amino-acid
- the biological environment
- the properties of the codon itself ( $XYZ$ ).

We neglect the time in which the biosynthesis process takes place as we assume that the biosynthesis processes are considered at the same time, at least compared to the time scale of evolution of the genetic code. We define the *branching ratio*  $B_{ZV}$  as

$$B_{ZV} = \frac{P(XYZ \rightarrow \text{a.a.})}{P(XYV \rightarrow \text{a.a.})} \quad (8)$$

We 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  $J_\alpha, J_\alpha^3$ , where  $\alpha = H$  or  $V$ , and on an other set of quantum labels leaving out the degeneracy on  $J_\alpha$ ; in Table 1 different irreducible representations with the same values of  $J_\alpha$  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.:

$$B_{ZV} = F_{ZV}(b.o.; IR(XYZ); IR(XYV)) \quad (9)$$

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.

We consider the quartets and sextets. There are five quartets and three sextets in the eukariotic code: that will allow a rather detailed analysis. Moreover the 3 sextets appear as the sum of a quartet and a doublet, see Table 1. In the following we consider only the quartet sub-part of the sextets. 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.

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

$$B'_{ZV} = F_{ZV}(b.o.; IR(X'Y'Z); IR(X'Y'V)) \quad (10)$$

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

$$B_{ZV} = B'_{ZV} \quad (11)$$

The analysis was performed on a set of data retrieved from the data bank of “Codon usage tabulated from GenBank” [3]. In particular we analyzed two different data set: the first one comprises all the data of at least 2 000 codons, while the second set represents all the data with at least 30 000 codons. The referring organism for the analysis was *Homo sapiens*, whose codon usage table derives from the analysis of more than 12 500 coding sequences, and corresponds to about 6 000 000 codons.

Three quartets, coding the amino-acids Pro, Ala and Thr, have exactly the same content in irreducible representations, see Table 1. In Table 2 we report the 16 biological organisms with highest statistics. In Figs. 2, 3 and 4 the  $B_{AG}$ ,  $B_{UG}$  and  $B_{CG}$  are reported for the 8 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 (8)) 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  $(1/2, 1/2)^1$  for the whole set. The relative position of each of these quartets of codons can be more easily visualized in Fig. 1 where Pro, Ala, Thr and Ser (quartet part) constitute the four edges of a vertical column linking the representation  $(1/2, 1/2)^1$ , sitting at the ground floor, first to the representation  $(3/2, 1/2)^1$ , then to the  $(1/2, 3/2)^1$  one and finally to the representation  $(3/2, 3/2)$ , this last one located at the top floor.
- 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 (8) 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  $(1/2, 1/2)^3$ . Considering Fig. 1, it is now the two representations  $(1/2, 1/2)^3$  and  $(1/2, 3/2)^2$  which are brought together, the codons associated to Val and Leu (quartet part) determining the vertices of two parallel and vertical plaquettes.
- 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. Indeed we can note in Fig. 1 that the representations  $(1/2, 1/2)^2$  and  $(3/2, 1/2)^2$  are connected by the codon quartet relative to Arg and (but) only by this multiplet. We also remark the Gly quartet in the representations  $(1/2, 3/2)^1$  and  $(3/2, 3/2)$ : its position is completely different from the above discussed quartets which show up in these representations.

Then in Figs. 5, 6 and 7 we have drawn the normalized branching ratios  $\hat{B}_{PG}$ ,  $P \in \{A, U, C\}$ , defined by:

$$\hat{B}_{PG} = \frac{B_{PG}}{\sum_{a.a.} B_{AG}} \quad (12)$$

where the sum  $\sum_{a.a.}$  is extended to the eight amino-acids above listed. The mean value and the standard deviation are:

|                                | Pro  | Ala  | Thr  | Ser  | Val  | Leu  | Arg  | Gly  |
|--------------------------------|------|------|------|------|------|------|------|------|
| $\langle \hat{B}_{AG} \rangle$ | 1.60 | 1.46 | 1.57 | 1.61 | 0.16 | 0.11 | 0.50 | 1.00 |
| $\sigma(\hat{B}_{AG})$         | 0.16 | 0.16 | 0.17 | 0.21 | 0.03 | 0.02 | 0.15 | 0.29 |
| $\langle \hat{B}_{CG} \rangle$ | 1.24 | 1.66 | 1.46 | 1.83 | 0.26 | 0.23 | 0.60 | 0.73 |
| $\sigma(\hat{B}_{CG})$         | 0.15 | 0.18 | 0.15 | 0.23 | 0.05 | 0.04 | 0.16 | 0.18 |
| $\langle \hat{B}_{UG} \rangle$ | 1.49 | 1.71 | 1.24 | 2.07 | 0.25 | 0.19 | 0.45 | 0.60 |
| $\sigma(\hat{B}_{UG})$         | 0.26 | 0.13 | 0.14 | 0.32 | 0.06 | 0.04 | 0.22 | 0.22 |

These diagrams show an universal behaviour of  $\hat{B}_{PG}$  which has the same value independently of the biological organism. We have omitted in the diagram the branching ratio of the amino-acid Gly as it is dependent from the branching ratios of the other amino-acids due to our definition eq. (12). In our model this behaviour can easily be understood if the branching ratio  $B_{ZV}$  has the factorized form

$$B_{ZV} = \Phi_{ZV}(b.o.) \psi_{ZV}(IR(XYZ); IR(XYV)) \quad (13)$$

This factorization explains also the correlation in the behaviour between the values of  $B_{PG}$  for different biological organisms, see Figs. 2, 3 and 4. Finally we report in the table below the mean value and the standard deviation for the case of biological organisms with low statistics to put in evidence the effects of the statistics.

|                                | Pro  | Ala  | Thr  | Ser  | Val  | Leu  | Arg  | Gly  |
|--------------------------------|------|------|------|------|------|------|------|------|
| $\langle \hat{B}_{AG} \rangle$ | 1.77 | 1.49 | 1.67 | 1.13 | 0.17 | 0.15 | 0.61 | 1.01 |
| $\sigma(\hat{B}_{AG})$         | 0.67 | 0.40 | 0.49 | 0.56 | 0.09 | 0.18 | 0.34 | 0.44 |
| $\langle \hat{B}_{CG} \rangle$ | 1.29 | 1.55 | 1.52 | 1.82 | 0.25 | 0.23 | 0.62 | 0.71 |
| $\sigma(\hat{B}_{CG})$         | 0.47 | 0.39 | 0.41 | 0.53 | 0.08 | 0.08 | 0.32 | 0.32 |
| $\langle \hat{B}_{UG} \rangle$ | 1.50 | 1.61 | 1.26 | 2.09 | 0.25 | 0.19 | 0.51 | 0.60 |
| $\sigma(\hat{B}_{UG})$         | 0.58 | 0.39 | 0.39 | 0.64 | 0.10 | 0.09 | 0.28 | 0.32 |

## 4 Conclusions

The basic elements of our model of the genetic code are the 4 nucleotides and the 64 codons come out as composed states. The symmetry algebra  $\mathcal{U}_{q \rightarrow 0}(sl(2) \oplus sl(2))$  has two main characteristics. Firstly, it encodes the stereochemical property of a base conferring quantum numbers to each nucleotide. Secondly, it admits representation spaces with the remarkable property that the vector bases of the tensor product are ordered sequences of the basic elements (nucleotides). 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. Indeed, as we have shown in this paper, it may explain the correlation between the branching ratio of the codon usage of different codons coding the same amino-acid. Let us remark that the assignments of the codons to the different irreducible representations is a straightforward consequence of the tensor product once assigned the nucleotides to the fundamental irreducible representation, see our first assumption.

It is a prevision of our model that for *any biological organism* belonging to the vertebrate series, in the limit of large number of biosynthesized amino-acids, the ratios  $B_{AG}$ ,  $B_{UG}$  and  $B_{CG}$  for, respectively, Pro, Ala, Thr and Ser (Val and Leu) should be 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. Our analysis has also shown an universal behaviour of the normalized branching ratio of the codon usage for the vertebrates, which was not evidently expected in our model, but which can easily be explained assuming a factorized form for the  $B_{ZV}$ . So, assuming the factorization (13), we foresee that the normalized ratio  $\widehat{B}_{AG}$ ,  $\widehat{B}_{UG}$  and  $\widehat{B}_{CG}$  should be given for any biological organism by the values reported in Figs. 5, 6 and 7.

A first analysis including biological organisms belonging to the invertebrate and plant series show that the pattern of correlation is still present, even in a less striking way, but significant deviations appear for some biological organisms. A more detailed analysis with extension to the other multiplets, in particular the doublets, and to other series of biological organisms will be done in a further more detailed publication.

**Acknowledgments** We are deeply indebted with Maria Luisa Chiusano for providing us the data which have allowed the analysis presented in this work and for very useful discussions. It is also a pleasure to thank J.C. Le Guillou for discussions and encouragements.

## References

- [1] L. Frappat, A. Sciarrino, P. Sorba, *A crystal basis for the genetic code*, Preprint ENSLAPP-AL-671/97 and DSF-97/37, physics/9801027, to appear in Phys. Lett. A.
- [2] M. Kashiwara, Commun. Math. Phys. **133** (1990) 249.
- [3] Y. Nakamura, T. Gojobori, and T. Ikemura, Nucleic Acids Research **26** (1998) 334.



Figure 1: Classification of the codons in the different crystal bases.

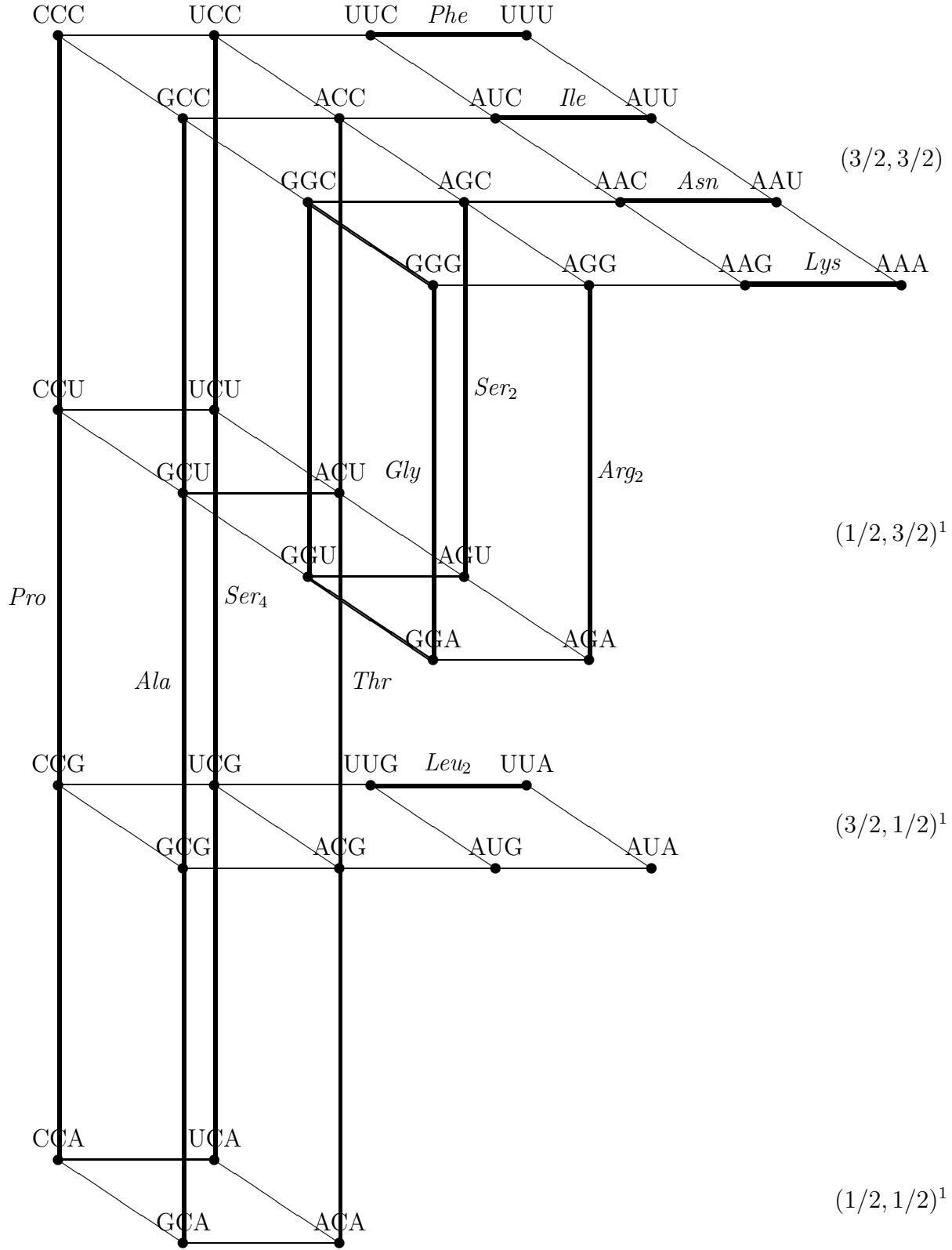


Figure 1 (cont'd)

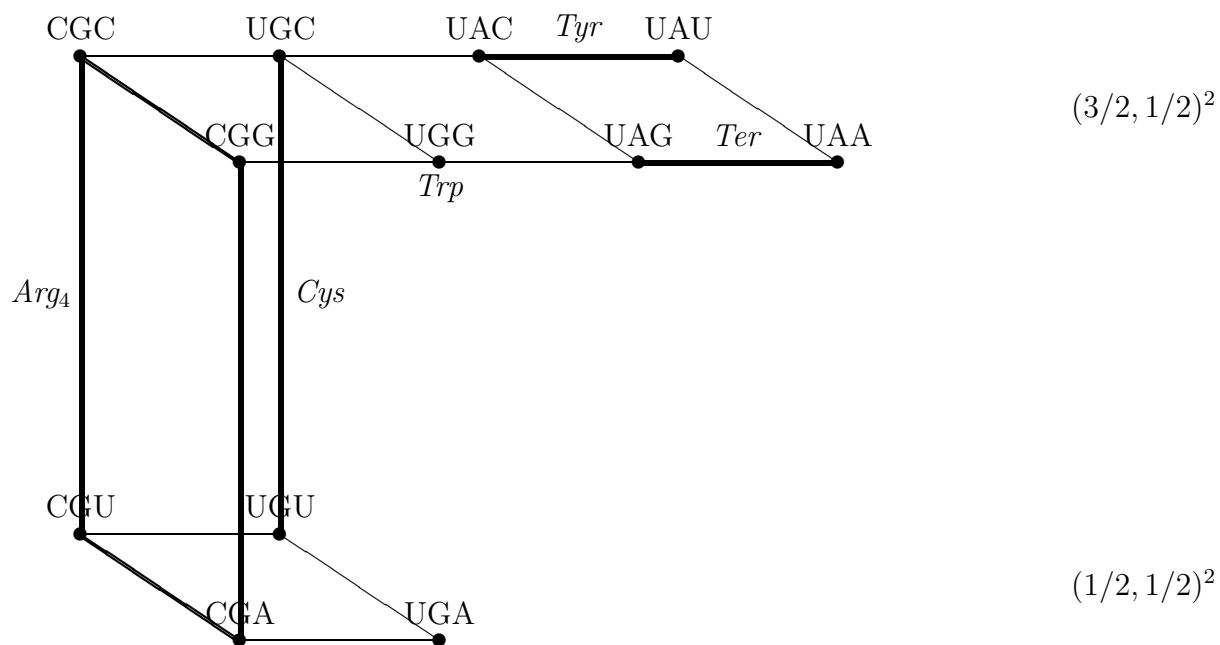
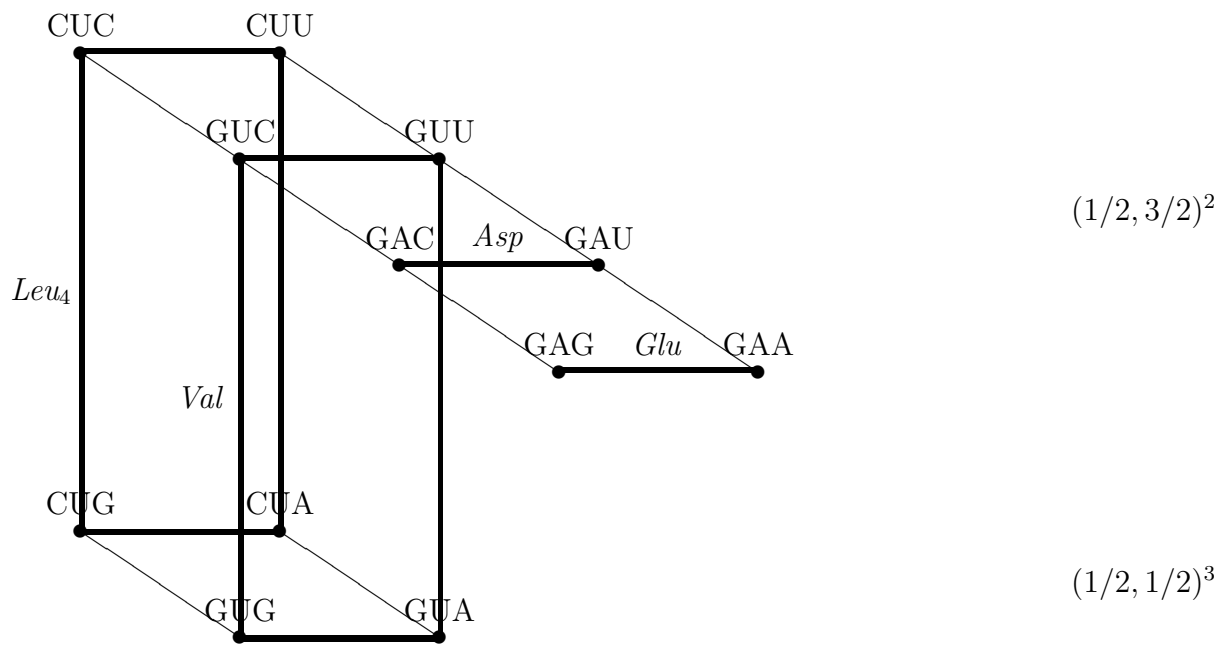


Figure 2: Branching ratio  $B_{AG}$  for the vertebrate series.

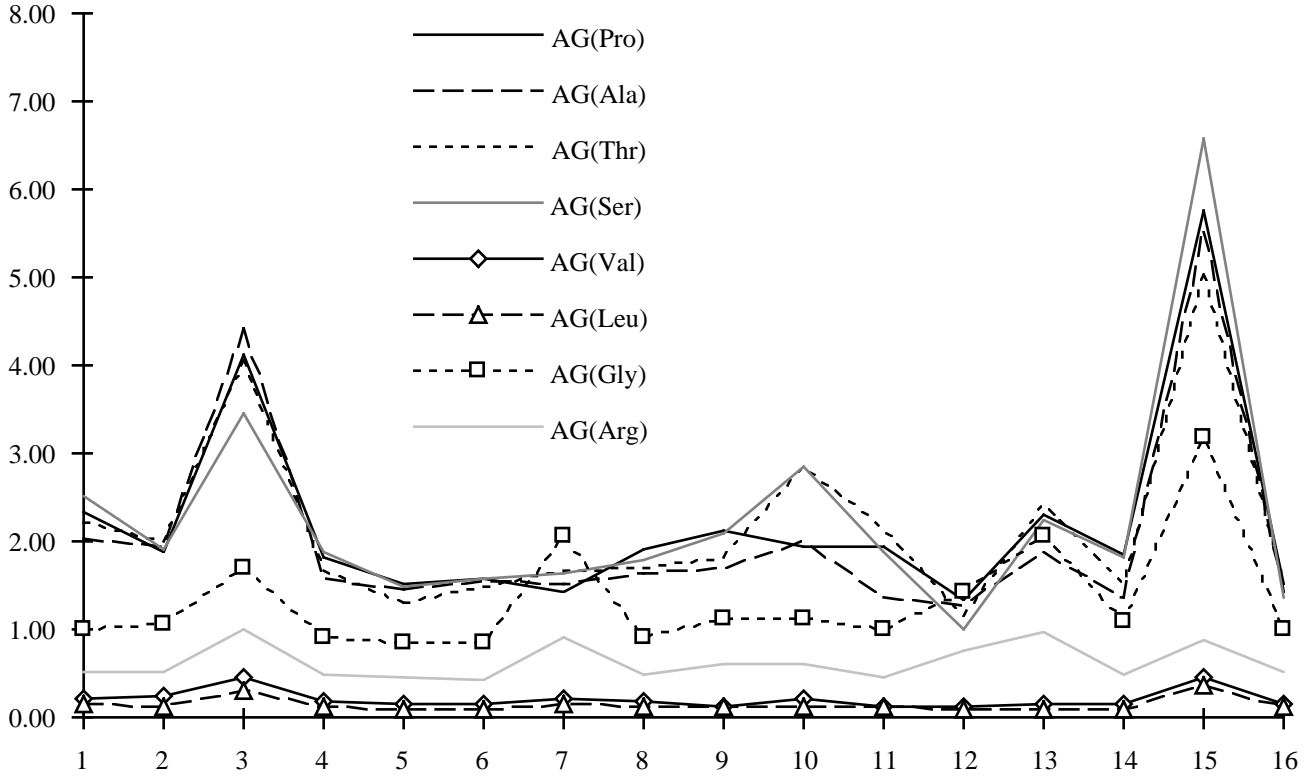


Figure 3: Branching ratio  $B_{CG}$  for the vertebrate series.

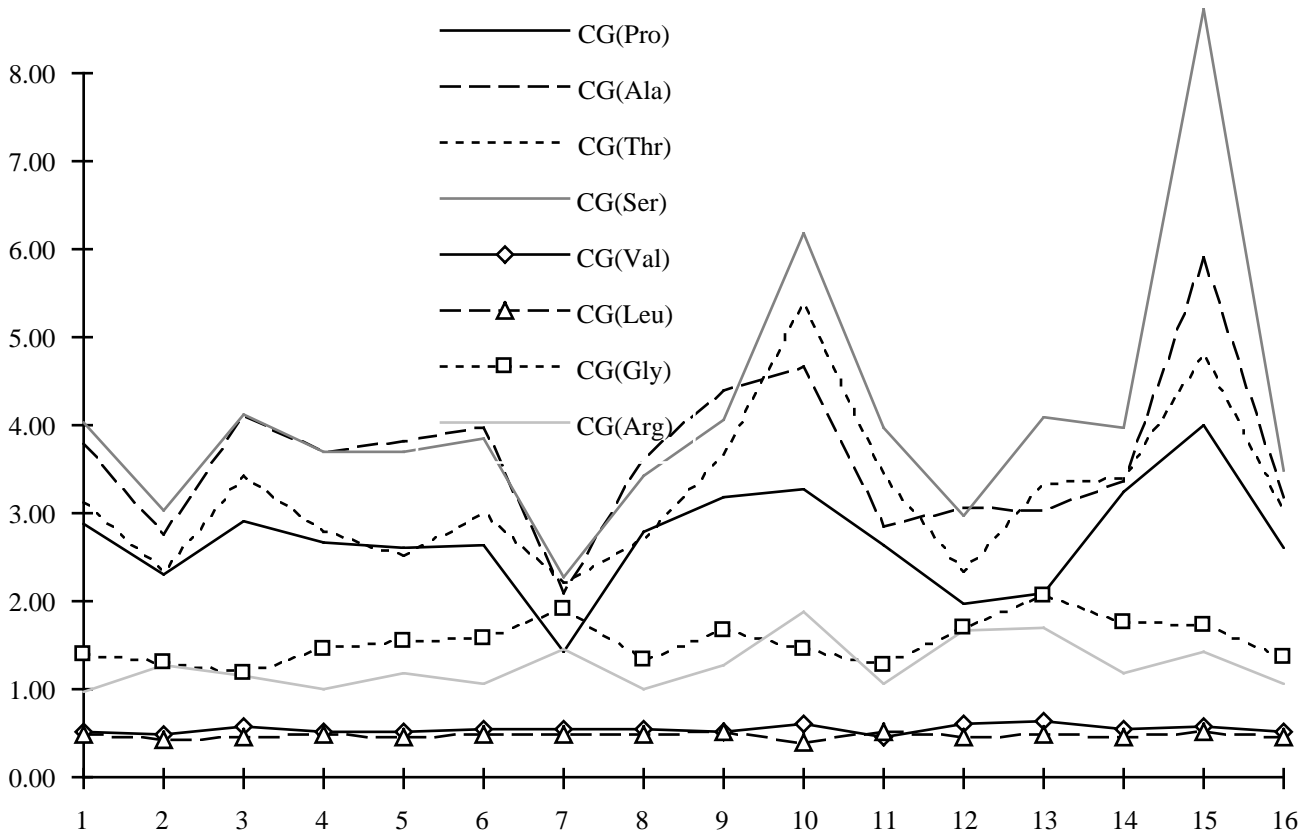


Figure 4: Branching ratio  $B_{UG}$  for the vertebrate series.

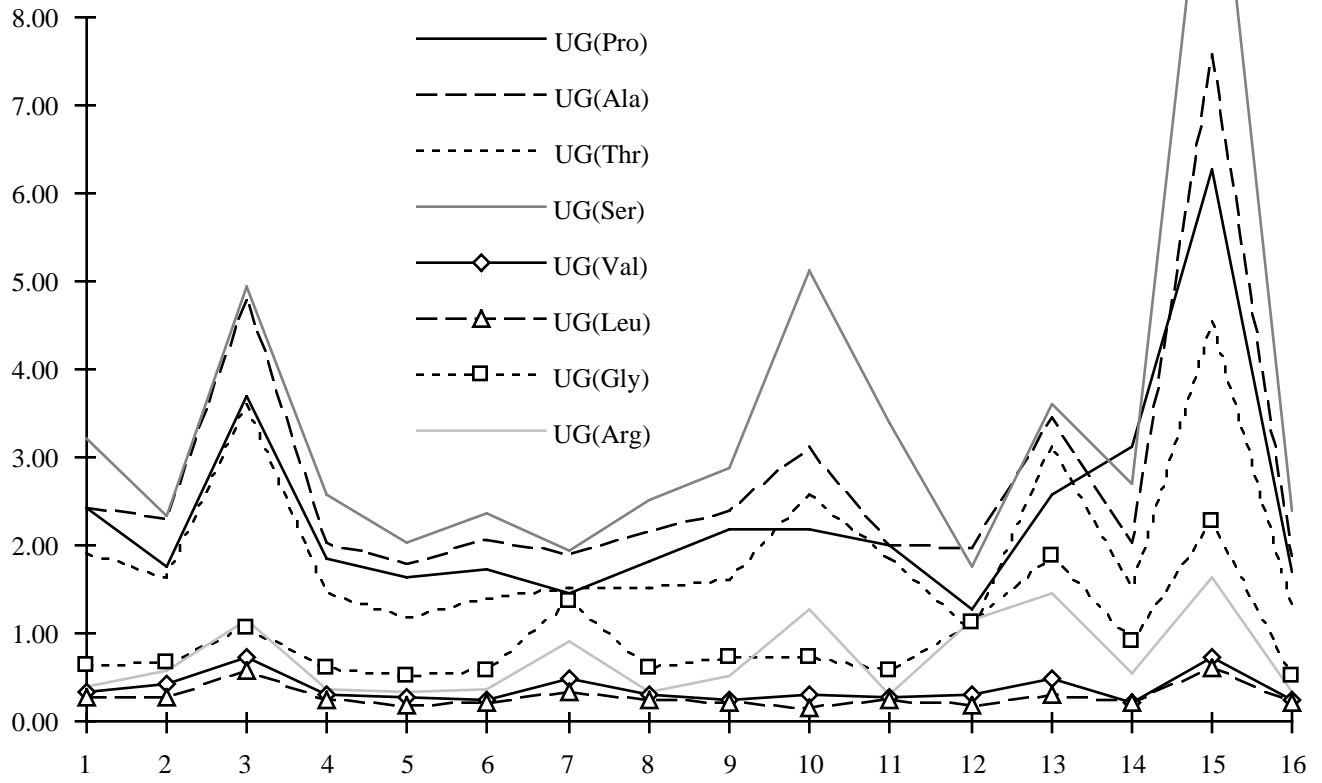


Figure 5: Normalized branching ratio  $B_{AG}$  for the vertebrate series.

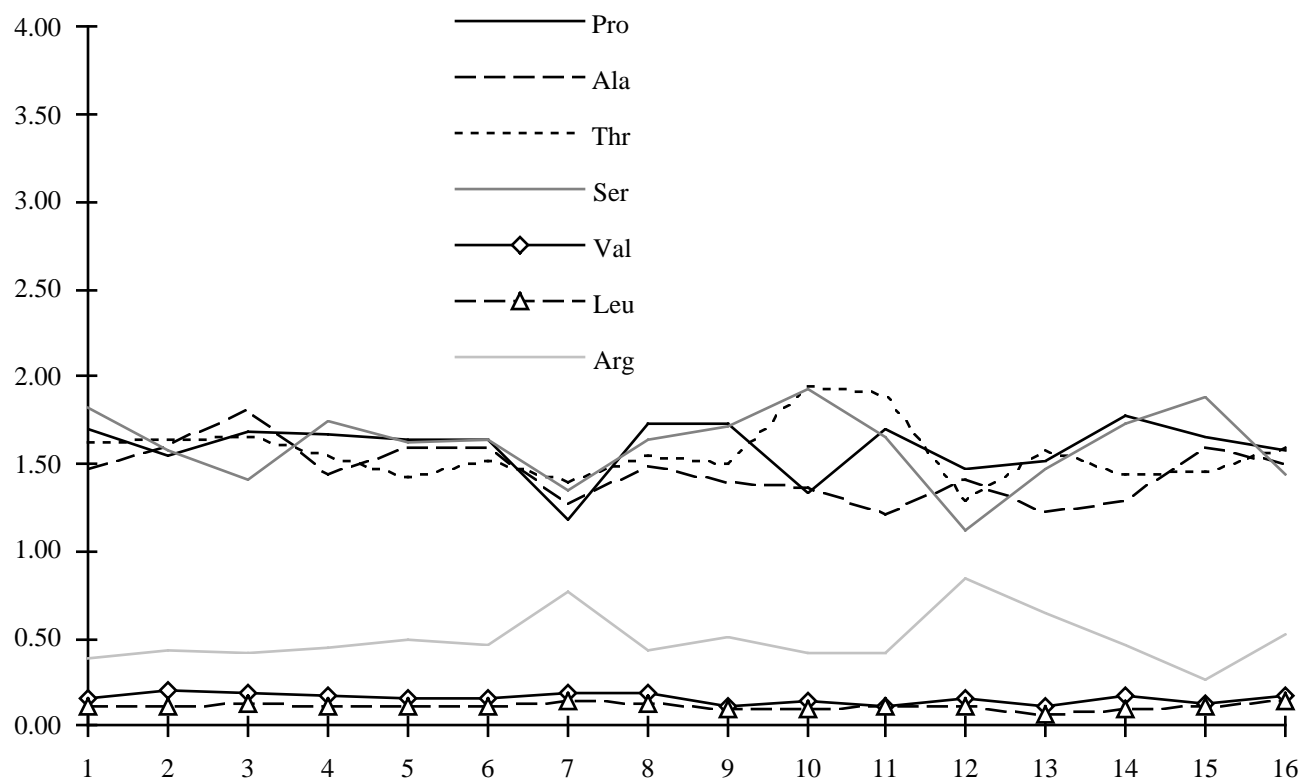


Figure 6: Normalized branching ratio  $B_{CG}$  for the vertebrate series.

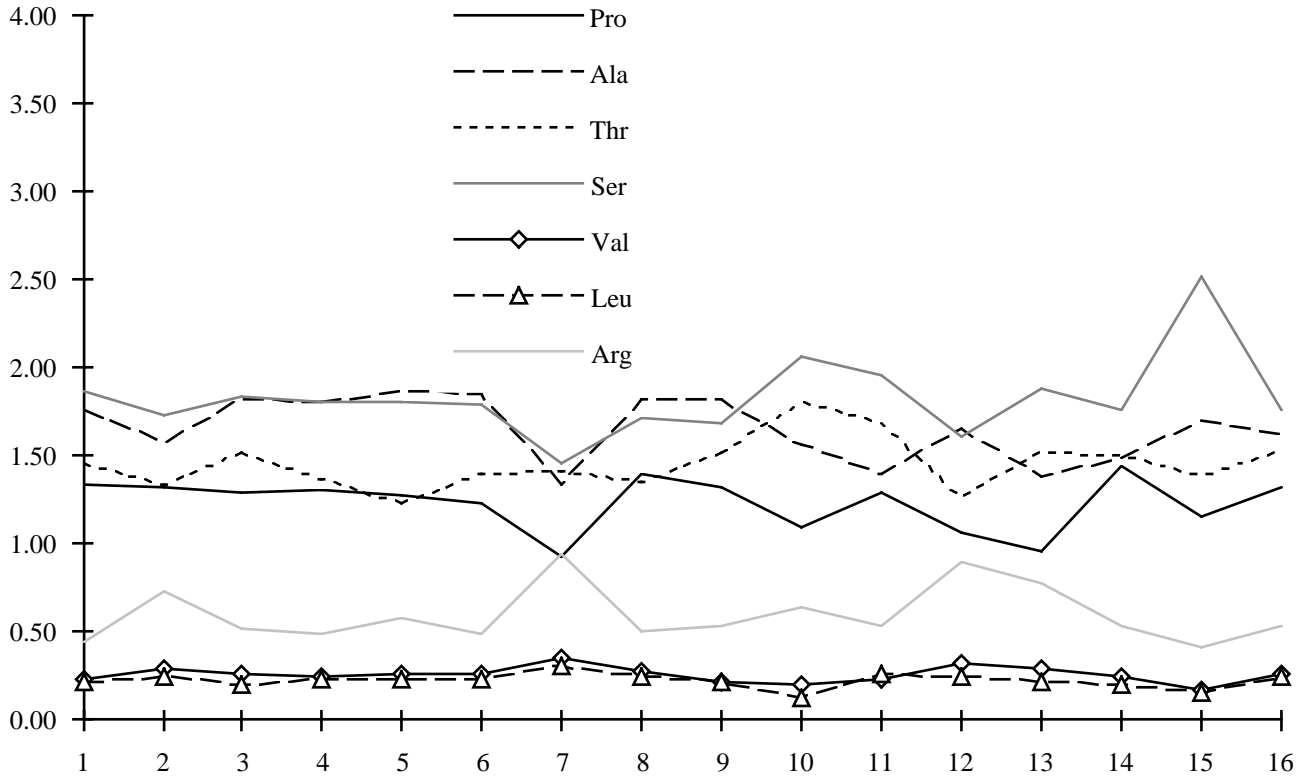


Figure 7: Normalized branching ratio  $B_{UG}$  for the vertebrate series.

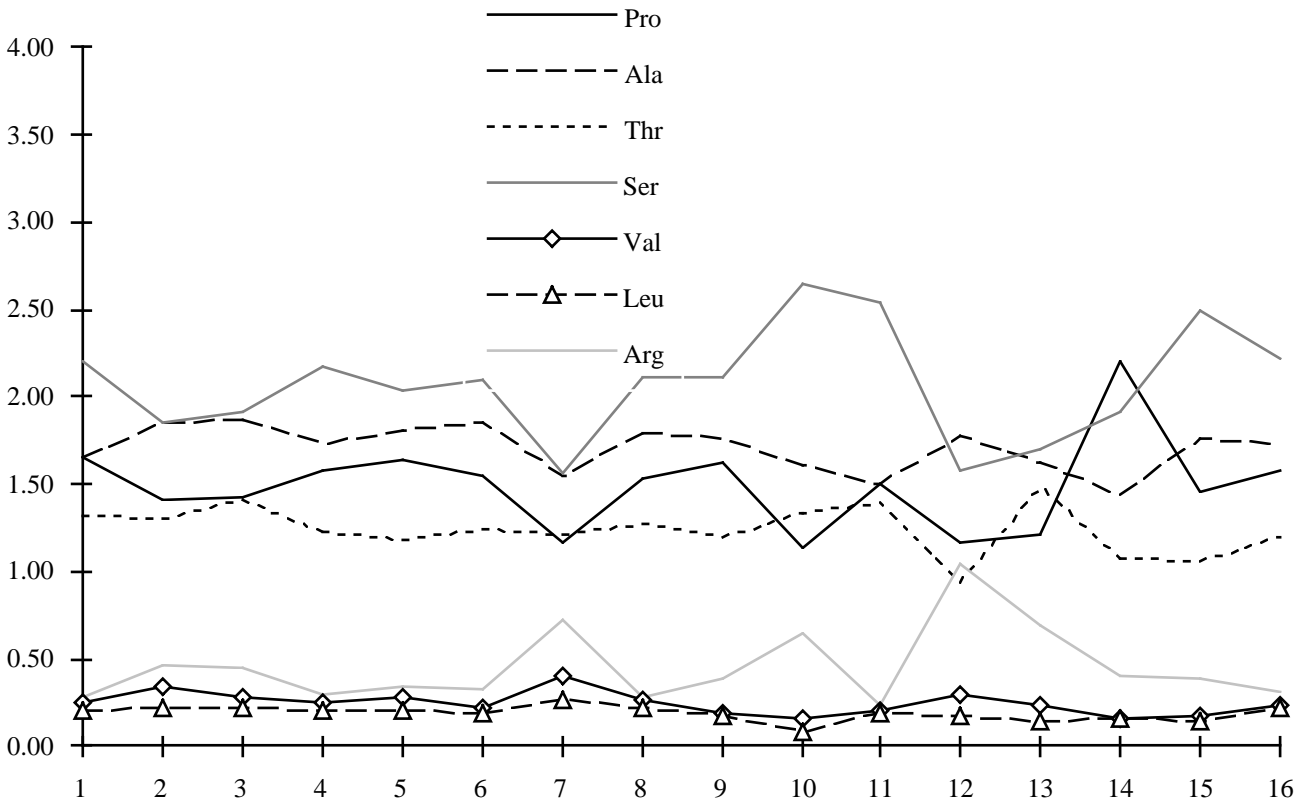


Table 1: The eukariotic code. The upper label denotes different IR.

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

Table 2: Biological organisms with highest statistics.

|    | Biological organism   | number of sequences | number of codons |
|----|-----------------------|---------------------|------------------|
| 1  | Homo sapiens          | 12 512              | 6 130 940        |
| 2  | Gallus gallus         | 1 319               | 638 532          |
| 3  | Xenopus laevis        | 1 144               | 493 437          |
| 4  | Bos taurus            | 1 182               | 478 270          |
| 5  | Oryctolagus cuniculus | 639                 | 321 129          |
| 6  | Sus scrofa            | 539                 | 216 654          |
| 7  | Danio rerio           | 259                 | 99 766           |
| 8  | Canis familiaris      | 230                 | 94 444           |
| 9  | Ovis aries            | 275                 | 81 177           |
| 10 | Oncorhynchus mykiss   | 128                 | 42 794           |
| 11 | Macaca mulatta        | 110                 | 34 510           |
| 12 | Fugu rubripes         | 63                  | 32 943           |
| 13 | Cyprinus carpio       | 95                  | 32 365           |
| 14 | Equus caballus        | 94                  | 31 254           |
| 15 | Rana cates beiana     | 61                  | 30 629           |
| 16 | Felis catus           | 83                  | 30 031           |