About a Symmetry of the Genetic Code

A. J. Koch and J. Lehmann

Institut de physique Expérimentale, Université de Lausanne, CH-1015 Lausanne, Switzerland

(*Received on* 22 *May* 1997, *Accepted on* 24 *June* 1997)

Considering the three codonic positions as independent, it is possible to explain the grouping of the 64 trinucleotides (without *AAA*, *TTT*, *CCC* and *GGG*) into three equal sets T_0, T_1 and T_2 according to their preferable reading frame (0, 1 or 2) in coding sequences. Supposing that the two complementarity strands of DNA are coding, it is demonstrated that the complementary of a codon is classified into the same set, which has been observed statistically (with a few exceptions) in the coding sequences by Arquès & Michel [(1996) *J. theor. Biol.* **182,** 45–58] and Arquès *et al.* [(1997) *J. theor. Biol.* **185,** 241–253]. Finally, the circular property of the code pointed out by these authors is demonstrated, and a direct consequence to biological considerations of these properties is discussed.

The *Journal of Theoretical Biology* recently published two articles concerning an extraordinary symmetry linked to codon frequencies discovered by Arquès & Michel (1996), and Arquès *et al.* (1997). After a short summary of their results, we would like to show that simple probabilistic considerations give some insight to their observations. This could help to understand the emergence of such symmetries in the code and could be of some interest to other readers who have, like us, been impressed by the mentioned articles.

Consider the 60 triplets obtained by removing from the 64 trinucleotides of the genetic code the four triplets *AAA*, *TTT*, *CCC* and *GGG*. For any nucleotide α, its complementary is noted $C(\alpha)$ (e.g. $C(A) = T$); similarly for a trinucleotide $\alpha\beta\gamma$, one defines $C(\alpha\beta\gamma) = C(\gamma)C(\beta)C(\alpha)$. At last, a circular permutation of a triplet is defined as $\mathscr{P}(\alpha\beta\gamma) = \beta\gamma\alpha$.

The trinucleotides are then grouped in three sets X_0, X_1 and X_2 so as to verify the following properties:

(i) each of the three sets X_0, X_1 and X_2 contains 20 trinucleotides.
(ii) $\mathscr{P}(X_0) = X_1$, $\mathscr{P}(X_1) = X_2$ and $\mathscr{P}(X_2) = X_0$ (*circular permutation property*).
(iii) $C(X_0) = X_0$, $C(X_1) = X_2$ and $C(X_2) = X_1$ (*complementarity property*).
(iv) $C(X_0)$, $C(X_1)$ and $C(X_2)$ are *maximal circular codes*.

The characteristics (i–iv) do not define uniquely the three sets and there are hundreds of possibilities to group the trinucleotides in three such sets.

Analysing the EMBL Nucleotide Sequence Data Library, Arquès & Michel (1996) grouped the trinucleotides of coding DNA sequences into three sets T_0, T_1 and T_2 according to the following rules: all triplets whose occurrence is higher in the normal reading frame (frame 0) than in frames shifted by one or two nucleotides in the 5′–3′ direction are grouped in T_0; the trinucleotides occurring preferentially in a frame shifted by one nucleotide in the 5′–3′ direction (frame 1) are placed in T_1 and the triplets whose frequency is maximal in a frame shifted by two nucleotides (frame 2) are collected in T_2. Surprisingly, if one removes the four trinucleotides *AAA*, *TTT*, *CCC* and *GGG*, the experimentally determined sets T_0, T_1 and T_2 verify nearly exactly the four properties listed above. We have checked Arquès & Michel's results on the 44th release of the EMBL database for prokaryotes and found very similar results (these authors found two misclassified trinucleotides, *GTG* and *TGG*; in our results *GTG*, *TGG*, *GCA*, *TGC* and *TCT* were misclassified). These facts are impressive, especially with regard to their statistical

0022–5193/97/220171 + 04 $25.00/0/jt970503

nature. The question raises whether it is possible to find some simple rules to understand the observed facts.

Let us suppose that the probability $p_i(\alpha)$ of occurrence of a given base α at position $i (i \in \{1, 2, 3\})$ in a trinucleotide observed in a DNA strand read in frame 0 only depends on i (in other words, there are no correlations between successive bases on a DNA strand; the frequency of a base only depends on its position i in a trinucleotide read in frame 0). For historical (e.g. RNY fossil code) or for transcriptional reasons, the $p_i(\alpha)$ are position dependent; this can be checked by measuring these probabilities on protein coding DNA sequences (see Table 1). The probability of finding the triplet $\alpha\beta\gamma$ in the frame 0—the reading frame—is then given by $p_1(\alpha)p_2(\beta)p_3(\gamma)$. Suppose that $\alpha\beta\gamma$ belongs to T_0. By definition of T_0, this means

$$\alpha\beta\gamma \in T_0 \Leftrightarrow \{p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\gamma)p_2(\alpha)p_3(\beta) \text{ and } p_1(\alpha)p_2(\beta)p_3(\gamma) > p_1(\beta)p_2(\gamma)p_3(\alpha)\}. \quad (1)$$

One can then show without difficulty that $\mathscr{P}(\alpha\beta\gamma) = \beta\gamma\alpha$ belongs to T_1. The proof is the following (just remember that, for T_1, the frame is shifted by one nucleotide):

$$\beta\gamma\alpha \in T_1 \Leftrightarrow \{p_2(\beta)p_3(\gamma)p_1(\alpha) > p_2(\alpha)p_3(\beta)p_1(\gamma) \text{ and } p_2(\beta)p_3(\gamma)p_1(\alpha) > p_2(\gamma)p_3(\alpha)p_1(\beta)\}. \quad (2)$$

One sees immediately that (2) is equivalent to (1). Similarly, one shows that $\beta\gamma\alpha \in T_1 \Leftrightarrow \mathscr{P}(\beta\gamma\alpha) = \gamma\alpha\beta \in T_2$. As a consequences, in a DNA strand without correlations between successive bases, the 60 retained trinucleotides are naturally classified in three sets T_0, T_1 and T_2 verifying the two properties (i) and (ii). With the frequencies of Table 1, one obtains

$$T_0 = \{AAT, AAC, ATT, ATC, ACT, CAC, CTT, CTC, GAA, GAT, GAC, GAG, GTA, GTT, GTC, GTG, GCA, GCT, GCC, GCG\}$$

which contains 13 trinucleotides of Arquès & Michel's T_0 set.

TABLE 1

By analysing the 44th release of the EMBL database for prokaryotes, we obtained the nucleotide frequencies $p_i(\alpha)$ at position $i \in \{1, 2, 3\}$ of the reading frame. By measuring these databases for eukaryotes, one obtains very similar results. These frequencies have been obtained by treating 11 748 coding sequences containing about 12 007 000 nucleotides (we only retained complete coding sequences beginning with ATG)

Base α	$p_1(\alpha)$	$p_2(\alpha)$	$p_3(\alpha)$
A	0.276	0.315	0.222
T	0.166	0.285	0.268
C	0.204	0.228	0.268
G	0.354	0.172	0.242

$$\begin{array}{ccccccccc} & & p_1 & & p_2 & & p_3 & & \\ 5' & — & \alpha & — & \beta & — & \gamma & — & 3' \\ 3' & — & C(\alpha) & — & C(\beta) & — & C(\gamma) & — & 5' \\ & & q_3 & & q_2 & & q_1 & & \end{array}$$

FIG. 1. The two complementary strands of a DNA molecule are both supposed to code in paired frames 0. Due to the complementarity of the two strands, the probabilities $p_i(\alpha)$ and $q_i(\alpha)$ to find the nucleotide α at position i in a triplet on the first or on the second strand are related (see text).

To reproduce requirement (iii), we have to add a hypothesis to the previous one concerning the $p_i(\alpha)$. Such an extra assumption is suggested by a remark from Arquès & Michel (1996). They write: "*As the set X_0 of trinucleotides is a circular code (. . .) and self-complementary, the two paired frames 0 (reading frames) in the two DNA double helix (sic) may simultaneously code for amino acids without using a start codon, in agreement with biological arguments* . . .".

Let us exploit this idea. Suppose that the two strands simultaneously code in paired frames 0. What does this imply at the level of trinucleotide frequencies? Consider Fig. 1. As previously, we write $p_i(\alpha)$ for the probability to find the nucleotide α at position i of the reading frame on the first DNA strand. On the paired strand, we note $q_i(\alpha)$ for the probability to find α at position i in frame 0. Since the two strands are paired, we have $p_1(\alpha) = q_3(C(\alpha))$ and $p_2(\alpha) = q_2(C(\alpha))$. Now, if both strands are equivalent for the coding function, our hypothesis of nucleotide frequencies linked to position in the reading frame implies that $p_i(\alpha) = q_i(\alpha)$. Combining these relations, one finds

$$p_1(\alpha) = p_3(C(\alpha)) \quad \text{and} \quad p_2(\alpha) = p_2(C(\alpha)). \quad (3).$$

These equalities are not precisely verified in Table 1. It is however not our purpose to explain these deviations here. By use of (3), one proves that $\alpha\beta\gamma \in T_0$ implies $C(\alpha\beta\gamma) \in T_0$. Suppose that $\alpha\beta\gamma \in T_0$; relations (l) are valid; now $C(\alpha\beta\gamma) \in T_0$ means

$$p_1(C(\gamma))p_2(C(\beta))p_3(C(\alpha)) > p_1(C(\beta))p_2(C(\alpha))p_3(C(\gamma))$$

and

$$p_1(C(\gamma))p_2(C(\beta))p_3(C(\alpha)) > p_1(C(\alpha))p_2(C(\gamma))p_3(C(\beta)).$$

By introducing (3) into the preceding inequalities, one gets

$$C(\alpha\beta\gamma) \in T_0 \Leftrightarrow \{p_3(\gamma)p_2(\beta)p_1(\alpha) > p_3(\beta)p_2(\alpha)p_1(\gamma) \text{ and } p_3(\gamma)p_2(\beta)p_1(\alpha) > p_3(\alpha)p_2(\gamma)p_1(\beta)\}$$

which is exactly (1). By using the very same procedure, one proves without difficulty that $\alpha\beta\gamma \in T_1 \Leftrightarrow C(\alpha\beta\gamma) \in T_2$. Within this theoretical framework, condition (iii) is a consequence of the equivalence of the two DNA strands concerning their coding functions.

The whole reasoning presented here relies on the fact that there are no correlations between successive bases, which is certainly incorrect. To check the validity of this hypothesis, we have measured Shannon's conditional information (Schmitt *et al.*, 1993; Lio *et al.*, 1996; Chatzidimitriou-Dreismann *et al.*, 1996) on the genes found in the EMBL database for prokaryotes; we have compared this result to the Shannon entropy measured on a random file constructed so as to verify the observed base frequencies at the three codon positions as listed in Table 1.

For words containing one to six bases, the ratio r between conditional information measured on the biological sequences and the random ones is close to $1{:}r > 0.995$ (data not shown here). The conclusion of this analysis is that, in a first approximation, correlations can be neglected since the conditional information remains nearly constant for short words and that the information measured on the biological data and on the random file are nearly the same for short words.

Let us finally show that T_0 (as well as T_1 and T_2) is a circular code. This means that any "circular" word w formed by concatenating trinucleotides of T_0 and written on a circle (the word is a circular concatenation of trinucleotides) is decomposable into series of elements of T_0 in a unique way.

To show this, we need properties (i) and (ii). Consider $w = \alpha_1\beta_1\gamma_1\alpha_2\beta_2\ldots\gamma_n$ $(n \geqslant 1)$ with $\alpha_k\beta_k\gamma_k \in T_0$ $(k = 1, \ldots n)$; by definition of T_0, we have [see eqn (1)]:

$$p_1(\alpha_k)p_2(\beta_k)p_3(\gamma_k) > p_1(\beta_k)p_2(\gamma_k)p_3(\alpha_k).$$

By construction of w, one has

$$\prod_{k=1}^{n} p_1(\alpha_k)p_2(\beta_k)p_3(\gamma_k) > \prod_{k=1}^{n} p_1(\beta_k)p_2(\gamma_k)p_3(\alpha_k). \quad (4)$$

Let us decompose w in the shifted frame 1: $w = \beta_1\gamma_1\alpha_2\beta_2\ldots\gamma_n\alpha_1$. Suppose that, in frame 1, w can be decomposed into elements of T_0, i.e. that $\beta_k\gamma_k\alpha_{k+1} \in T_0$ $(k = 1\ldots n)$ (with the obvious convention $\alpha_{n+1} = \alpha_1$). This implies that

$$p_1(\beta_k)p_2(\gamma_k)p_3(\alpha_{k+1}) > p_1(\alpha_{k+1})p_2(\beta_k)p_3(\gamma_k),$$

and, by construction of w:

$$\prod_{k=1}^{n} p_1(\beta_k)p_2(\gamma_k)p_3(\alpha_{k+1}) = \prod_{k=1}^{n} p_1(\beta_k)p_2(\gamma_k)p_3(\alpha_k) >$$

$$> \prod_{k=1}^{n} p_1(\alpha_{k+1})p_2(\beta_k)p_3(\gamma_k) = \prod_{k=1}^{n} p_1(\alpha_k)p_2(\beta_k)p_3(\gamma_k).$$

This relation is in contradiction with (4). As a consequence, w cannot be decomposed on T_0 in frame 1 (remember however the statistical nature of this result).

To show that w cannot be decomposed on T_0 in frame 2, one proceeds similarly. In conclusion, the decomposition of w on T_0 is unique so that T_0 is a circular code. It is even a maximal circular code, since it contains 20 elements. In a similar way, one demonstrates that T_1 and T_2 are also maximal circular codes.

At this stage of the analysis, let us mention a biological consequence of the complementarity property (iii). An important parameter to characterize an organic molecule is its hydrophobicity. If one ranks the nucleic acids from most hydrophobic to most hydrophilic (Jungck, 1978), the succession is

$$A\ G\ C\ U.$$

Blalock and Smith (1984) [see also Taylor and Coates (1989)] have shown that the hydrophobicity of an amino acid is strongly correlated to the one at the mid-base position of the associated codon(s): a hydrophobic amino acid nearly always has a U at the mid-base position of its associated codon(s); a hydrophilic one very often has an A at the same position. As a consequence, given a coding DNA sequence, it is possible to estimate the hydrophobicity of the encoded protein.

In Fig. 2(a), we have calculated the number of U at the second position of the triplets in each protein vs the number of A at the same position for 11456 proteins. One clearly recognizes two classes of proteins characterized by a different ratio U/A at position 2. No such difference appears for the ratio G/C. We separate arbitrarily the two classes by putting the limit at $U/A = 1.43$.

Using the hydrophobicity ranking obtained by Lacey *et al.* (1983) for the amino acids, we crudely estimate the average hydrophobicity rank of a protein in the following way:

average hydrophobicity rank of the protein = sum of the hydrophobicity ranks of aa in protein/number of aa.

A low rank corresponds to a high value of hydrophobicity and vice versa. By plotting the average hydrophobicity ranks of the proteins [Fig. 2(b)], one notices an astonishing correlation between the protein hydrophobicity and the two classes defined by the ratio U/A at position 2:

class I $U/A > 1.43$ protein with low hydrophobicity rank,

class II $U/A \leqslant 1.43$ protein with high hydrophobicity rank.

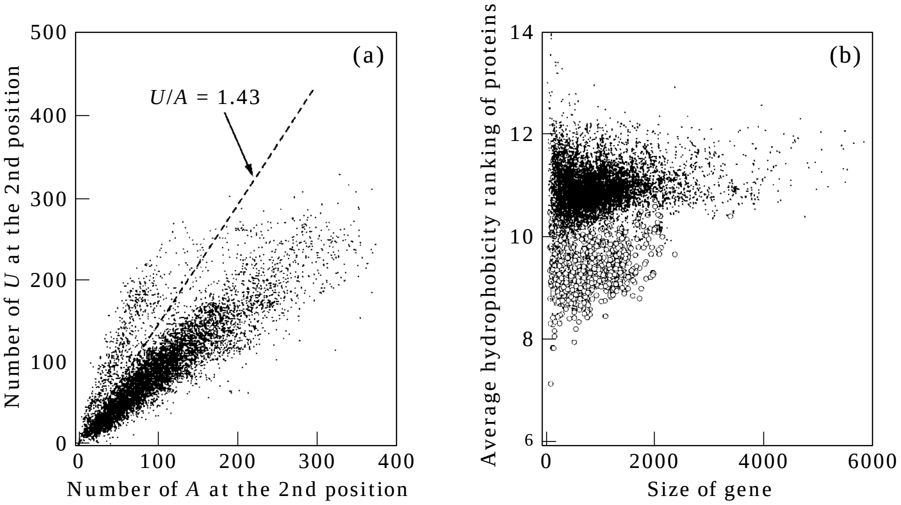

FIG. 2. (a) Number of A vs number of U at the second position of the codon. Each point corresponds to one gene. Genes with $U/A > 1.43$ are assigned to class I; genes with $U/A \leqslant 1.43$ are assigned to class II (see text). (b) Hydrophobicity of proteins plotted vs the size of their corresponding genes (size of genes is reported to spread out horizontally the data and has no significance here); genes belonging to class I are marked by a circle; genes of class II are marked by a dot. Each dot or circle corresponds to one gene. 11456 prokaryotic genes have been analysed and reported.

There is a clear partition of the proteins according to the ratio U/A at codon position 2. It is probable that class I groups together most of the transmembrane proteins because they always contain hydrophobic domains, while class II collects the intracellular proteins. To check the validity of these results, the same investigations were made for organelle proteins. One finds the same correspondence between hydrophobicity and classes of U/A but with a higher proportion of genes in class I; since organelles have proportionally more transmembrane proteins, this result is reliable.

To return to considerations about the property (iii), let us point out that the complementary bases A and U are at opposite extremities of the hydrophobic succession list. Now, if one imagines two strands of coding DNA with paired frames 0 (Fig. 1), one has automatically position 2 (which is discriminant for hydrophobicity) complementary to position 2 on the other strand. If the two paired DNA strands eventually both code for proteins in paired frames 0 as suggested by the complementarity property (iii), then one can assert that the two proteins produced by the DNA strands will exhibit opposite hydrophobicities (relatively to a mid-value).

The statistical symmetry discovered by Arquès & Michel can be interpreted as resulting from the inhomogeneous distribution of the four bases when measured in a DNA strand on the three positions of the reading frame. These frequencies allow the grouping of the trinucleotides into three sets T_0, T_1 and T_2 verifying properties (i) and (ii), by assuming that there are no correlations between successive bases. Furthermore, the resulting sets show property (iv) to be maximal circular codes.

By assuming, as suggested by these authors, that the two strands of the DNA molecule are equivalent for the coding function in frame 0 one obtains further the complementarity property (iii).

Within the framework developed here, the few exceptions to this statistical symmetry observed in biological data need to be explained by the existence of correlations between nucleotides on a DNA strand.

REFERENCES

ARQUÈS, D. G. & MICHEL, C. J. (1996). A complementary circular code in the protein coding genes. *J. theor. Biol.* **182,** 45–58.

ARQUÈS, D. G., FALLOT, J.-P. & MICHEL, C. J. (1997). An evolutionary model of a complementary circular code. *J. theor. Biol.* **185,** 241–253.

BLALOCK, J. E. & SMITH, E. M. (1984). Hydropathic anti-complementarity of amino-acids based on the genetic code. *Biochem. Biophys. Rev. Commun.* **121,** 203–207.

CHATZIDIMITRIOU-DREISMAMM, C. A., STERIFFER, R. M. F. & LARHAMMAR, D. (1996). Lack of biological significance in the linguistic features of non-coding DNA—a quantitative analysis. *Nucl. Acids Res.* **24,** 1676–1681.

JUNGCK, J. R. (1978). The genetic code as a periodic table. *J. mol. Evol.* **11,** 211–224.

LACEY, J. C. & MULLINS, D. W. JR. (1983). Experimental studies related to the origin of the genetic code and the process of protein synthesis—a review. *Origins of Life* **13,** 3–42.

LIO, P., POLITI, A., BUIATTI, M. & RUFFO, S. (1996). High statistics block entropy measures of DNA sequences. *J. theor. Biol.* **180,** 151–160.

SCHMITT, A. O., HERZEL, H. & EBELING, W. (1993). A new method to calulate higher-order entropies from finite samples. *Europhys. Lett.* **23,** 303–309.

TAYLOR, F. J. R. & COATES, D. (1989). The code within the codons. BioSystems **22,** 177–187.