J. theor. Biol. (1994) **167**, 413–414

LETTER TO THE EDITOR

On the Origin of the Periodicity of Three in Protein Coding DNA Sequences

Since the early observation by Shepherd (1981*a*, *b*) it is well known that most mature messenger RNAs have a rhythm with a period of three bases. The phenomenon has been associated to some statistical properties of protein-coding DNA, such as the distribution of purines (R) and pyrimidines (Y), and the predominance of RNY codons (N = any base) (Shepherd, 1981*a*, *b*; Arques & Michel, 1992), codon usage bias (Rowe & Trainor, 1983) and abundance of codons with the same base at first or second codon position (Tsonis *et al.*, 1991). Constraints related to mechanism that monitors the translation frame have also been invoked (Trifonov, 1987; Lagúnez-Otero & Trifonov, 1992).

In this letter, we would like to recall that the periodicity of three in protein-coding DNA sequences appears primarily as a consequence of the non-uniform distribution of base composition at the codon sites, a well known phenomenon which has allowed the design of algorithms to detect protein coding regions in a DNA sequence (see Staden, 1990 for a review). Such heterogeneity can be measured, for example, by the variance of each base by codon position.

To substantiate our insight we shall use two examples. The first concerns the human apolipoprotein gene (13 692 bases), whose nucleotide sequence shows a clear periodicity of three which was attributed by Tsonis *et al.* (1991) to the major 35·8% of codons having A at codon position II. These authors suggest that, in general, the periodicity of three in protein-coding DNA sequences is originated by the predominant presence of codons having the same base at positions I or II.

The frequencies and variances of each base within the apolipoprotein gene are shown in Table 1(a), and the spectra of the four numerical series obtained by seriatim labelling of each base as 1 and the others as 0 are plotted on Fig. 1(a). By comparing plots and data, a clear parallelism between the heterogeneity of frequencies of each base at the three codon sites and the height of the peak at period three can be seen. In other words, the peak for base G (which is rather scarce at codon position II) is the highest due to the strong compositional differences at the three codon sites, while the peak for base C, which is more evenly distributed at the three codon sites, is the lowest. The abundance of base A at codon position II is only a secondary factor as will be shown below.

The second example concerns a computer generated random sequence (15 000 bases) whose compositional features are given in Table 1(b) and the spectra are plotted on Fig. 1(b). This sequence has two bases (A and C) with the same variances and different frequencies, and two bases (G and T) with the same frequencies and different variances. Looking at the graphs, it can be seen that: (i) When the variances are equal (A and C), the highest peak corresponds to the base with higher frequency (C), and (ii) when the frequencies are equal (G and T), the highest peak corresponds to the base with greater variance (G).

The secondary role of the relative frequencies of any base on the periodicity of three is demonstrated, since there is no correspondence between the order of any of the major base frequencies by position

TABLE 1

Base frequencies by codon site and variances (S) *of human apoliprotein gene* (*a*) *and a computer random-generated sequence* (*b*)

	(a)				(b)			
	A	C	G	T	A	C	G	T
I	32·5	21·2	27·8	18·5	25·0	45·0	11·0	19·0
II	35·8	22·1	12·0	30·1	16·0	35·0	24·0	25·0
III	24·8	25·3	21·6	28·3	05·0	26·0	40·0	30·0
Total	31·0	22·9	20·5	25·6	15·3	35·6	24·4	24·7
$S(\times 10^3)$	3·1	0·4	6·4	3·8	10·0	9·0	21·0	3·0

0022–5193/94/080413 + 02 $8.00/0

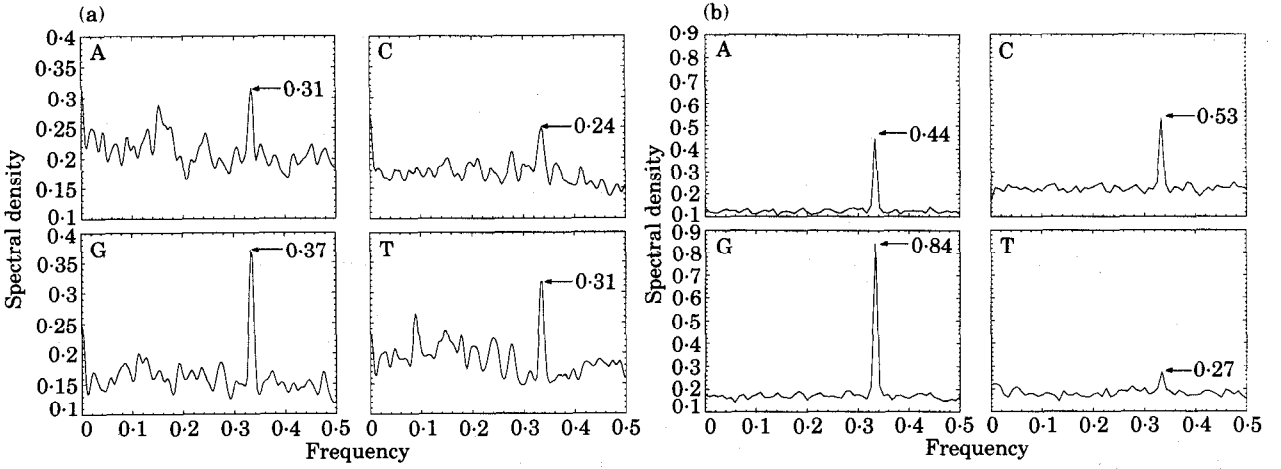

FIG. 1. Spectral densities of bases A, C, G and T of (a) human apolipoprotein gene and (b) of a computer-generated random sequence. The height of the peaks at period three (frequency = 0·333) are indicated by arrows.

(C at I, G at III, T at III, and A at I) and the order of the heights of the peaks (G, C, A and T).

GABRIEL GUTIÉRREZ,† JOSÉ L. OLIVER‡ AND ANTONIO MARÍN§‖

† § *Departamento de Genética, Facultad de Biología, Universidad de Sevilla, Apartado* 1095, *E*-41080 *Sevilla, Spain,*
‡ *Departamento de Genética, Facultad de Ciencias, Universidad de Granada, E*-18071 *Granada, Spain.*
(*Received on* 3 *August* 1993, *Accepted on* 6 *October* 1993)

REFERENCES

ARQUES, D. G. & MICHEL, C. J. (1992). *J. theor. Biol.* **156,** 113–127.
LAGÚNEZ-OTERO, J. & TRIFONOV, E. N. (1992). *J. Biomolec. Struct. Dyn.* **10,** 455–464.
ROWE, G. W. & TRAINOR, L. E. H. (1983). *J. theor. Biol.* **42,** 245–261.
SHEPHERD, J. C. W. (1981*a*). *J. molec. Evol.* **17,** 94–102.
SHEPHERD, J. C. W. (1981*b*). *Proc. natn. Acad. Sci. U.S.A.* **78,** 1596–1600.
STADEN, R. (1990). *Meth. Enzymol.* **183,** 163–180.
TRIFONOV, E. N. (1987). *J. molec. Biol.* **194,** 643–652.
TSONIS, A. A., ELSNER, J. B. & TSONIS, P. A. (1991) *J. theor. Biol.* **151,** 323–331.

‖ Author to whom correspondence should be addressed at: Departamento de Genética, Universidad de Sevilla, Apartado 1095, E-41080 Sevilla, Spain.