J. theor. Biol. (1990) **143**, 307-318

Periodicities in Coding and Noncoding Regions of the Genes

DIDIER G. ARQUÈS† AND CHRISTIAN J. MICHEL‡§

† *Université de Franche-Comté, Laboratoire d'Informatique de Besançon, Unité Associée CNRS No.* 822, 16, *Route de Gray*, 25030 *Besançon, France and* ‡ *Friedrich Miescher Institut, Bioinformatic group, Mattenstrasse* 22, *P.O. Box* 2543, *CH*-4002 *Basel, Switzerland*

(*Received on* 29 *July* 1989, *Accepted in revised form on* 30 *November* 1989)

Gene population statistical studies of protein coding genes and introns have identified two types of periodicities on the purine/pyrimidine alphabet: (i) the modulo 3 periodicity or coding periodicity (periodicity P3) in protein coding genes of eukaryotes, prokaryotes, viruses, chloroplasts, mitochondria, plasmids and in introns of viruses and mitochondria, and (ii) the modulo 2 periodicity (periodicity P2) in the eukaryotic introns. The periodicity study is herein extended to the 5′ and 3′ regions of eukaryotes, prokaryotes and viruses and shows: (i) the periodicity P3 in the 5′ and 3′ regions of prokaryotes and viruses, and (ii) the periodicities P2 and P3 in the 5′ and 3′ regions of eukaryotes. Therefore, these observations suggest a unitary and dynamic concept for the genes as for a given genome, the 5′ and 3′ regions have the genetic information for protein coding genes and for introns:

(1) In the eukaryotic genome, the 5′ (P2 and P3) and 3′ (P2 and P3) regions have the information for protein coding genes (P3) and for introns (P2). The intensity of P3 is high in 5′ regions and weak in 3′ regions, while the intensity of P2 is weak in 5′ regions and high in 3′ regions.
(2) In the prokaryotic genome, the 5′ (P3) and 3′ (P3) regions have the information for protein coding genes (P3).
(3) In the viral genome, the 5′ (P3) and 3′ (P3) regions have the information for protein coding genes (P3) and for introns (P3). The absence of P2 in viral introns (in opposition to eukaryotic introns) may be related to the absence of P2 in 5′ and 3′ regions of viruses.

1. Introduction

Statistical studies of gene populations are further investigated on the two letter alphabet $=\{R, Y\}$ (R = purine, Y = pyrimidine, $N = R$ or Y). Let a motif m be a concatenation of several letters (e.g. the motif YRY) and let an i-motif m_i be two identical motifs m separated by any i bases N and noted $m(N)_im$ [e.g. the i-motif $YRY(N)_iYRY$]. The occurrence study of the i-motif $YRY(N)_iYRY$ allows to analyse the 6-motif $YRY(N)_6YRY$ which may have a central function in the DNA sequence evolution (Arquès & Michel, 1987*b*) and also to reveal periodicities. Precisely, there is a periodicity Pr if for some i_0 in the range $[0, r-1]$, the i-motif $YRY(N)_iYRY$ has a preferential occurrence for i congruent to i_0 modulo r (noted $i \equiv i_0[r]$). Two periodicities (statistically defined in the method section) having important biological meanings, were identified in gene populations: the periodicity

§ Author to whom correspondence should be addressed.

0022-5193/90/070307+12 $03.00/0

P3, called coding periodicity (Shepherd, 1981; Fickett, 1982; Arquès & Michel, 1987*a,b,c*) which was attributed to the preferential use of the *RNY* codon (Eigen, 1971; Eigen & Schuster, 1978), and the periodicity P2 (Arquès & Michel, 1987*c*) which was related to the alternating purine/pyrimidine stretches. Herein, this study will identify periodicities in 5′ and 3′ regions of eukaryotes, prokaryotes and viruses and will lead to a unitary and dynamic concept for genes as for a given genome, the 5′ and 3′ regions have the genetic information for protein coding genes and for introns.

2. Method

2.1. STATISTICAL FUNCTION

The method is identical to the one developed previously by Arquès & Michel (1987*b*) whose outlines are presented below. Let F be one of the gene populations (see Table 1) obtained from the EMBL Nucleotide Sequence Data Library (release 18). A gene population incorporates all sequences having enough information for its classification, e.g. a sequence with unspecified bases or with an unmentioned taxonomic group, is excluded. These gene populations are characterized by their notation, their number of sequences and by their number of kilobases (kb) (see Table 1). The 5′ regions studied are located upstream from an open reading frame which starts with an initiator ATG codon. The 3′ regions studied start with a stop codon TAA, TAG or TGA. Other types of 5′ and 3′ regions as well as those of chloroplasts and mitochondria (not enough sequences available and too similar sequences), were excluded from this survey. The population F has $n(F)$ sequences. Let s be a sequence in F with a length $l(s)$. Let m_i be the i-motif $m_i = YRY(N)_iYRY$ by varying i in the range $[0, 99]$, i.e. two trinucleotides YRY separated by any i bases N. For each s of F, the counter $c_i(s)$ counts the occurrences of m_i in s. In order to count the m_i occurrences in the same conditions for all i, only the first $l(s)-104$ [$=l(s)-(99+6)+1$] bases of s are examined ($99+6$ is the maximal length of m_i). Then, the occurrence probability $o_i(s)$ of m_i for s, is equal to $c_i(s)/[l(s)-104]$,

TABLE 1

Gene populations

(1) 5′ region populations
—Eukaryotes, noted *N*5EUK (1808 sequences, 1268 kb)
—Prokaryotes, noted *N*5PRO (650 sequences, 335 kb)
—Viruses, noted *N*5VIR (290 sequences, 197 kb)
(2) Intron populations
—Eukaryotes, noted *I*EUK (1396 sequences, 1000 kb)
—Viruses, noted *I*VIR (60 sequences, 106 kb)
(3) 3′ region populations
—Eukaryotes, noted *N*3EUK (2614 sequences, 1634 kb)
—Prokaryotes, noted *N*3PRO (350 sequences, 215 kb)
—Viruses, noted *N*3VIR (301 sequences, 265 kb)

i.e. the ratio of the counter by the total number of current bases read. Then, the occurrence probability $p_i(F)$ of m_i for F, is equal to $[\sum_{s\in F} o_i(s)]/n(F)$. For each F, the statistical function $i \rightarrow p_i(F)$ by varying i, is represented as a curve $C(F)$. A minimal length of 200 bases for the sequences analysed was chosen in order to have a sufficient number of m_{99} occurrences to give a sense to their occurrence probabilities.

2.2. PERIODICITIES

The periodicities P2 and P3 are statistically defined as follows:

—Periodicity P2 in a range $[0, b]$ (Arquès & Michel, 1987*c*):
$p_i(F) > \max\{p_{i-1}(F), p_{i+1}(F)\}$, $i \in [0, b]$ and $i \equiv 1[2]$.

—Periodicity P3 in a range $[a, 96]$ (Shepherd, 1981; Fickett, 1982; Arquès & Michel, 1987*a,b,c*):
$p_i(F) > \max\{p_{i-1}(F), p_{i+1}(F)\}$, $i \in [a, 96]$ and $i \equiv 0[3]$.

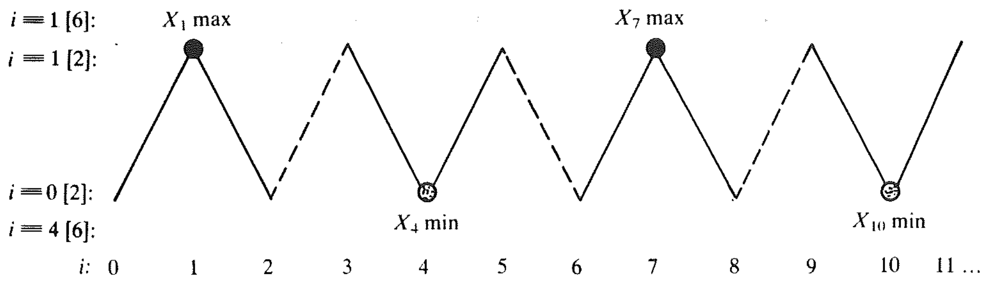
There is an incomplete periodicity P3 in a range $[a, 96]$ when a few points $[i, p_i(F)]$ do not satisfy the above inequality.

The periodicities P2 and P3 are identified and measured with Binomial tests (see below). Precisely, these tests are used to exclude a periodicity P2 (the presence of a periodicity P2 being obvious) and to demonstrate the presence of an incomplete periodicity P3.

2.2.1. *Periodicity P2*

For a population F with $i \equiv 1[6]$ (resp. with $i \equiv 4[6]$) in the range $[0, b]$, let $X_i(F)$ be the Bernoulli random variable which is equal to 1 with the probability p if $p_i(F) > \max\{p_{i-1}(F), p_{i+1}(F)\}$, i.e. a maximum or a peak [resp. if $p_i(F) < \min\{p_{i-1}(F), p_{i+1}(F)\}$, i.e. a minimum or a through], and 0 with the probability $1-p$ otherwise (see Table 2). The sum of the independent $X_i(F)$ is a Binomial random variable $M(F)$ of parameter p and of the order $n = \lfloor (b+2)/3 \rfloor (\lfloor \alpha \rfloor$ being

Table 2 Periodicity P2



the integer part of α). For example if $b = 11$ (see Table 2), $M(F) = X_1(F) + X_4(F) + X_7(F) + X_{10}(F)$ is a Binomial random variable, i.e. $M(F)$ is the sum of $n = \lfloor 13/3 \rfloor =$ four Bernoulli random variables $X_1(F)$, $X_4(F)$, $X_7(F)$ and $X_{10}(F)$ of parameter p. A curve $C(F)$ with the periodicity P2 is then associated to a parameter $p = 1$ (limit case), i.e. if $C(F)$ has the periodicity P2 then $X_i(F) = 1$ with the probability $1: X_i(F)$ is a Bernoulli random variable of parameter $p = 1$ and $M(F) = n$. To the contrary, a random (having no periodicity) curve $C(F)$ is associated to a parameter $p = 1/3$ [one chance out of three that a point $p_i(F)$ with $i \equiv 1[6]$ is higher than the two adjacent points and one chance out of three that a point $p_i(F)$ with $i \equiv 4[6]$ is lower than the two adjacent points].

2.2.2. *Periodicity P3*

For a population F with $i \equiv 0[3]$ in the range $[a, 96]$, let $Y_i(F)$ be the Bernoulli random variable which is equal to 1 with the probability p if $p_i(F) > \max\{p_{i-1}(F), p_{i+1}(F)\}$, and 0 with the probability $1-p$ otherwise. The sum of the independent $Y_i(F)$ is a Binomial random variable $N(F)$ of parameter p and of the order $n = 96/3 - \lceil a/3 \rceil + 1$ ($\lceil \alpha \rceil$ being the nearest greater integer of α), which counts the number of maxima (peaks) among the n possible values of i in the range $[a, 96]$. For example if $a = 80$, $N(F) = Y_{81}(F) + Y_{84}(F) + Y_{87}(F) + Y_{90}(F) + Y_{93}(F) + Y_{96}(F)$ is a Binomial random variable, i.e. $N(F)$ is the sum of $n = 96/3 - \lceil 80/3 \rceil + 1 = 6$ Bernoulli random variables $Y_{81}(F)$, $Y_{84}(F)$, $Y_{87}(F)$, $Y_{90}(F)$, $Y_{93}(F)$ and $Y_{96}(F)$ of parameter p. A curve $C(F)$ with a periodicity P3 (resp. incomplete periodicity P3) is associated to a parameter p equal (resp. close) to 1. To the contrary, a random curve $C(F)$ is associated to a parameter $p = 1/3$ (one chance out of three that a point $p_i(F)$ with $i \equiv 0[3]$ is higher than the two adjacent points). For a given population F, the hypothesis $H_0: p = 1/3$ is tested against the hypothesis $H_1: p > 1/3$ at the 1% level. If n is large enough [i.e. $n \times p \times (1-p) \geq 5$, i.e. $n \geq 22$], the central limit theorem asserts that under H_0, $Z(F) = (N(F) - \mu)/\sigma$ is close to a reduced centered Gaussian variable, $\mu = np$ and $\sigma = [np(1-p)]^{1/2}$ being the mean and the standard deviation of the Binomial distribution $N(F)$ respectively. Under $H_0: p = 1/3$, by replacing p by $1/3$ then $Z(F) = [N(F) - n \times 3^{-1}]/(n \times 2 \times 9^{-1})^{1/2}$ is identified with a reduced centred Gaussian variable. $H_0: p = 1/3$ is rejected and $H_1: p > 1/3$ is accepted if the variable $N(F)$ of mean value np is significantly greater than $n/3$ (mean value under H_0) and therefore if the variable $Z(F)$ is significantly greater than 0. $Z(F)$ being a reduced centred Gaussian variable and by choosing a statistical level of 1%, the table of the Gauss law shows that prob $[Z(F) > 2{\cdot}32] = 1\%$. Therefore, the hypothesis H_1 of the incomplete periodicity P3 is accepted at the 1% level if $Z(F) > 2{\cdot}32$.

3. Results

3.1. GENE POPULATIONS WITH THE PERIODICITY P2

In the range $[0, 23]$: *N*5EUK: Fig. 1(a).
—in the range $[0, 50]$: *I*EUK: Fig. 1(d).

—in the range [0, 99]: $N3EUK$: Fig. 1(f) [except for the points $[i, p_i(N3EUK)]$ at $i=35$, $i=41$, $i=43$ and $i=79$].

3.2. GENE POPULATIONS WITHOUT PERIODICITY P2 AND WITH THE INCOMPLETE PERIODICITY P3 IN THE RANGE [3, 96]

The gene populations $N5PRO$ [Fig. 1(b)], $N5VIR$ [Fig. 1(c)], $IVIR$ [Fig. 1(e)], $N3PRO$ [Fig. 1(g)] and $N3VIR$ [Fig. 1(h)] have no periodicity P2 in a range $[0, b]$. Indeed, if for example the range [0, 23] found with $N5EUK$ is chosen (then $n=8$) and by using the Binomial random variable $M(F)$ which characterizes the periodicity P2 (see section 2.2.1), then $M(F)=8$ if the curve $C(F)$ has the periodicity P2 and $M(F)=8/3$ (on average) if the curve $C(F)$ is random. The $M(F)$ values obtained with these five populations show that the curves $C(F)$ are random (i.e. absence of the periodicity P2):
—$N5PRO$: Fig. 1(b) $[M(N5PRO)=3,\ n=8]$.
—$N5VIR$: Fig. 1(c) $[M(N5VIR)=2,\ n=8]$.
—$IVIR$: Fig. 1(e) $[M(IVIR)=2,\ n=8]$.
—$N3PRO$: Fig. 1(g) $[M(N3PRO)=3,\ n=8]$.
—$N3VIR$: Fig. 1(h) $[M(N3VIR)=1,\ n=8]$.

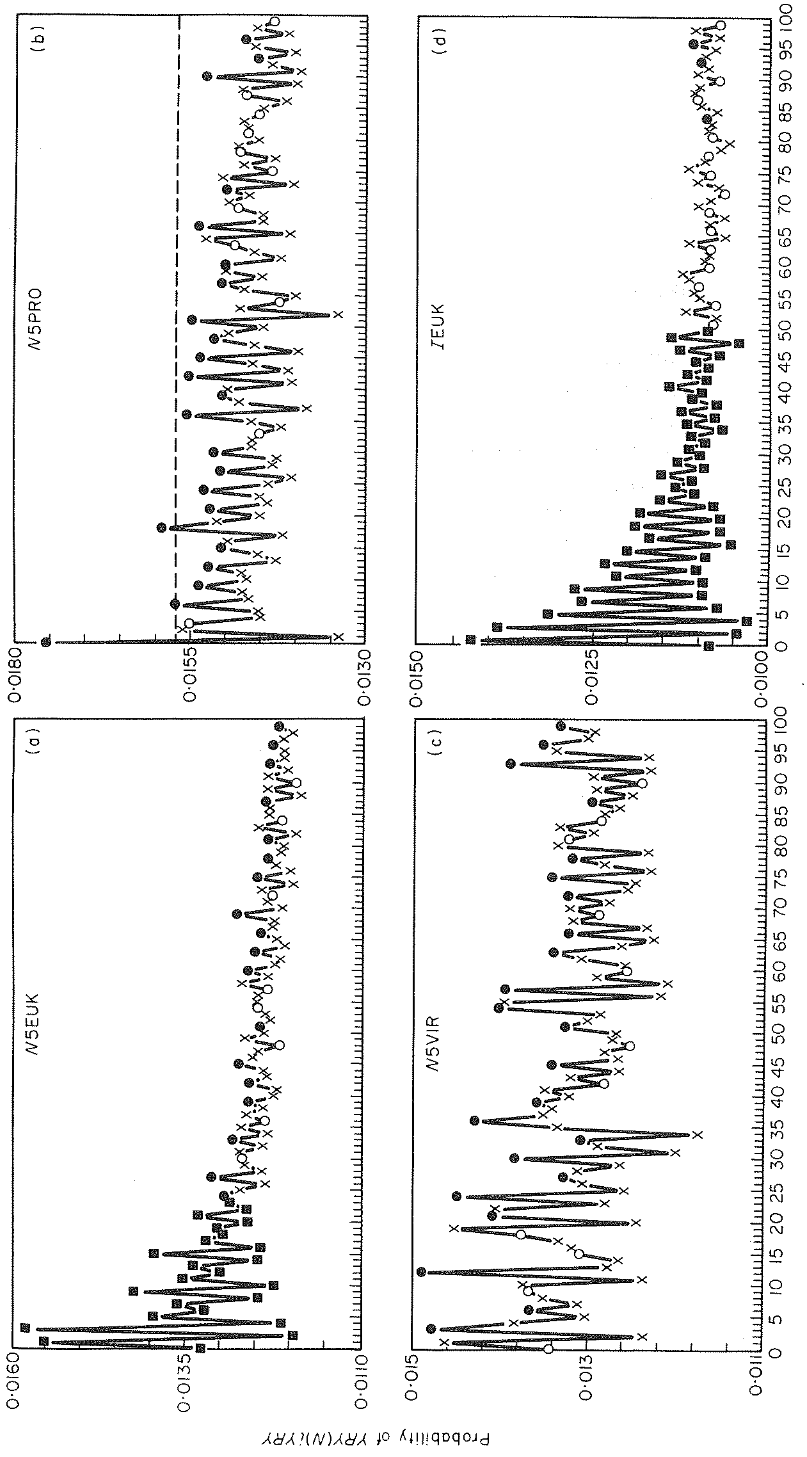
Since only two types of periodicities were significantly found in gene populations and since no periodicity P2 was identified in the populations $N5PRO$, $N5VIR$, $IVIR$, $N3PRO$ and $N3VIR$ (see above), then the hypothesis of a periodicity P3 can be tested against the random situation. In these five populations, the Binomial random variable $N(F)$ identifies an incomplete periodicity P3 (see section 2.2.2) in the range [3, 96] (then $n=32$) because $Z(F)>2{\cdot}32$:
—$N5PRO$: Fig. 1(b) $[N(N5PRO)=22,\ n=32$ and $Z(N5PRO)=4{\cdot}25]$.
—$N5VIR$: Fig. 1(c) $[N(N5VIR)=22,\ n=32$ and $Z(N5VIR)=4{\cdot}25]$.
—$IVIR$: Fig. 1(e) $[N(IVIR)=28,\ n=32$ and $Z(IVIR)=6{\cdot}50]$.
—$N3PRO$: Fig. 1(g) $[N(N3PRO)=22,\ n=32$ and $Z(N3PRO)=4{\cdot}25]$.
—$N3VIR$: Fig. 1(h) $[N(N3VIR)=19,\ n=32$ and $Z(N3VIR)=3{\cdot}12]$.

3.3. THE PERIODICITIES P2 AND P3 OCCUR SIMULTANEOUSLY IN THE GENE POPULATIONS $N5EUK$ AND $N3EUK$, BUT NOT IN $IEUK$

Surprisingly, a periodicity P3 after the periodicity P2 is identified in the population $N5EUK$. Indeed, $N5EUK$ [see Fig. 1(a)] has an incomplete periodicity P3 in the range [24, 96] $[N(N5EUK)=17,\ n=25$ and $Z(N5EUK)=3{\cdot}68]$. For $IEUK$ [see Fig. 1(d)], no periodicity P3 is observed in the range [51, 96] $[N(IEUK)=3,\ n=16]$.

The simultaneity of the periodicities P2 and P3 as well as the location of P3 after P2 in the population $N5EUK$ reveals two questions:

(1) It seems unlikely to have a "periodicity transformation" of P2 into P3 when i increases. Indeed, there is no simple model of motifs on the alphabet $=\{R, Y\}$ explaining such an observation. Therefore, this observation may be due to the simultaneity of a periodicity P3 in the total range [3, 96] with a strong periodicity P2 of decreasing intensity in the range [0, 23] which hides the weak periodicity P3


(a)
N5EUK
0·0160
0·0135
0·0110
(b)
N5PRO
0·0180
0·0155
0·0130
(c)
N5VIR
0·015
0·013
0·011
(d)
IEUK
0·0150
0·0125
0·0100
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
i
Probability of YRY(N)iYRY

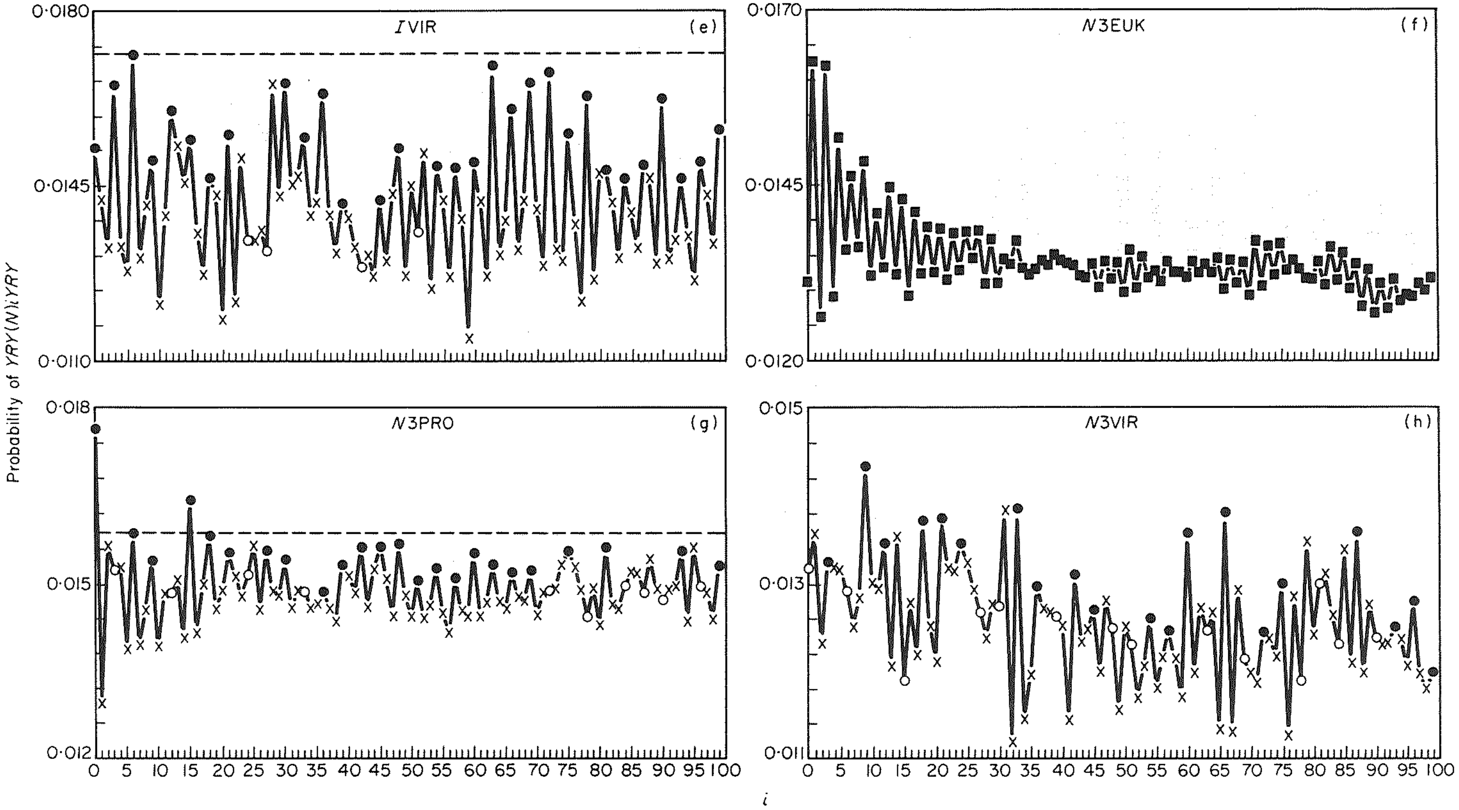


FIG. 1. Mean occurrence probability of the i-motif $YRY(N)_iYRY$ in gene populations. The horizontal axis represents the number i of bases N in the i-motif $YRY(N)_iYRY$, with i in the range $[0, 99]$. The vertical axis represents the mean occurrence probability $p_i(F)$ (see method) over all the sequences in the following gene populations: (a) *N*5EUK: 5′ eukaryotic regions, (b) *N*5PRO: 5′ prokaryotic regions, (c) *N*5VIR: 5′ viral regions, (d) *I*EUK: eukaryotic introns, (e) *I*VIR: viral introns, (f) *N*3EUK: 3′ eukaryotic regions, (g) *N*3PRO: 3′ prokaryotic regions, (h) *N*3VIR: 3′ viral regions. The points $[i, p_i(F)]$ are marked with black squares for a periodicity P2 and with circles and crosses for a periodicity P3: circles for the points with $i \equiv 0[3]$ [black circles if $p_i(F) > \max\{p_{i-1}(F), p_{i+1}(F)\}$] and crosses for the points with $i \equiv 1, 2[3]$. A horizontal dashed line goes through the point $[6, p_6(F)]$.

in the range [0, 23]. One way to prove this hypothesis is to suppress the large alternating purine/pyrimidine stretches which are known to be associated with the periodicity P2 (Arquès & Michel, 1987*c*) and to observe the periodicity P3 in the total range [3, 96].

Let F^* be the population issued from the population F in which the alternating purine/pyrimidine stretches of length > ten bases are suppressed [$2^{10}(=1024$ bases) is approximately the size of a sequence which allows a maximal length of ten bases for one stretch occurrence]. From the three populations having the periodicity P2 (*N*5EUK, *I*EUK and *N*3EUK), the three populations: *N*5EUK* (1808 sequences, 1248 kb), *I*EUK* (1396 sequences, 984 kb) and *N*3EUK* (2614 sequences, 1612 kb) are issued. *N*5EUK* [see Fig. 2(a)] has a periodicity P3 (complete) in the range [3, 21] and an incomplete periodicity P3 in the range [24, 96] [$N(N5\text{EUK}^*)=16$, $n=25$ and $Z(N5\text{EUK}^*)=3{\cdot}25$]. *I*EUK* has no periodicity P3 in the range [3, 96] (data not shown). Surprisingly, *N*3EUK* [see Fig. 2(b)] has an incomplete periodicity P3 in the range [3, 96] at the 2% statistical level [$N(N3\text{EUK}^*)=16$, $n=32$ and $Z(N3\text{EUK}^*)=2{\cdot}00$].

(2) The simultaneity of the periodicities P2 and P3 in the populations *N*5EUK and *N*3EUK should exist at the sequence level and should not be due to a partition of the population into two subpopulations, one with the periodicity P2 and the other with the periodicity P3. The concept of homogeneity of gene populations favours this hypothesis tested as follows:

Let $F_{10>}$ be the subpopulation of F incorporating only the sequences having alternating purine/pyrimidine stretches of length > ten bases. From the two populations *N*5EUK and *N*3EUK, the two subpopulations $N5\text{EUK}_{10>}$ (613 sequences, 636 kb) and $N3\text{EUK}_{10>}$ (804 sequences, 712 kb) are obtained. $N5\text{EUK}_{10>}$ [see Fig. 3(a)] has a periodicity P2 in the range [0, 30] and an incomplete periodicity P3 in the range [33, 96] [$N(N5\text{EUK}_{10>})=16$, $n=22$ and $Z(N5\text{EUK}_{10>})=3{\cdot}92$] (result similar to the *N*5EUK one). $N3\text{EUK}_{10>}$ [see Fig. 3(b)] has a periodicity P2 in the range [0, 99] except for the point $[i, p_i(N3\text{EUK}_{10>})]$ at $i=96$ (result similar to the *N*3EUK one). Furthermore, the suppression of the alternating purine/pyrimidine stretches of length > ten bases in $N5\text{EUK}_{10>}$ and $N3\text{EUK}_{10>}$ shows that the two subpopulations $N5\text{EUK}^*_{10>}$ (613 sequences, 616 kb) and $N3\text{EUK}^*_{10>}$ (804 sequences, 690 kb) have an incomplete periodicity P3 [see Fig. 4(a) and (b) respectively] in the range [3, 96] [$N(N5\text{EUK}^*_{10>})=24$, $n=32$, $Z(N5\text{EUK}^*_{10>})=5{\cdot}00$ and $N(N3\text{EUK}^*_{10>})=19$, $n=32$, $Z(N3\text{EUK}^*_{10>})=3{\cdot}12$ respectively] (results similar to the *N*5EUK* and *N*3EUK* ones).

The results found with the populations F, F^*, $F_{10>}$ and $F^*_{10>}$ are all in agreement with each other. Furthermore, all these results can be retrieved (data not shown) by applying this methodology to alternating purine/pyrimidine stretches of length less than ten bases (more sequences are concerned) or greater than ten bases (less sequences are concerned).

In conclusion, the 5′ and 3′ eukaryotic regions have both periodicities P2 and P3. The intensity of P3 is high in 5′ regions [$Z(N5\text{EUK})=3{\cdot}68$, $Z(N5\text{EUK}^*)=3{\cdot}25$, $Z(N5\text{EUK}_{10>})=3{\cdot}92$, $Z(N5\text{EUK}^*_{10>})=5{\cdot}00$] and weak in 3′ regions [no P3 in *N*3EUK and $N3\text{EUK}_{10>}$, $Z(N3\text{EUK}^*)=2{\cdot}00$, $Z(N3\text{EUK}^*_{10>})=3{\cdot}12$] while the

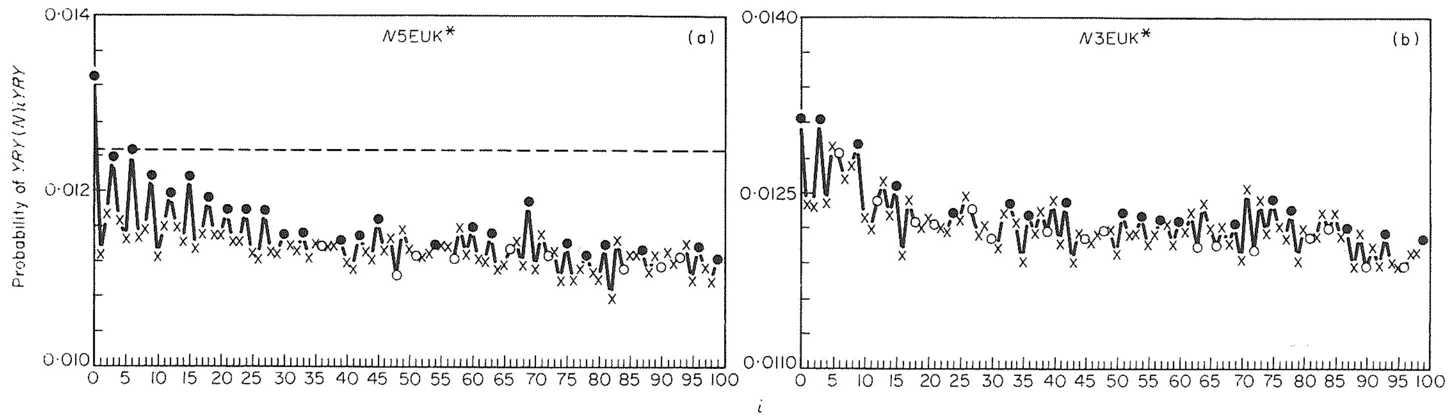


FIG. 2. Mean occurrence probability of the i-motif $YRY(N)_iYRY$ in the 5′ and 3′ eukaryotic region populations in which the alternating purine/pyrimidine stretches of length >10 bases are deleted (see method and legend in Fig. 1): (a) $N5\mathrm{EUK}^*$, (b) $N3\mathrm{EUK}^*$.

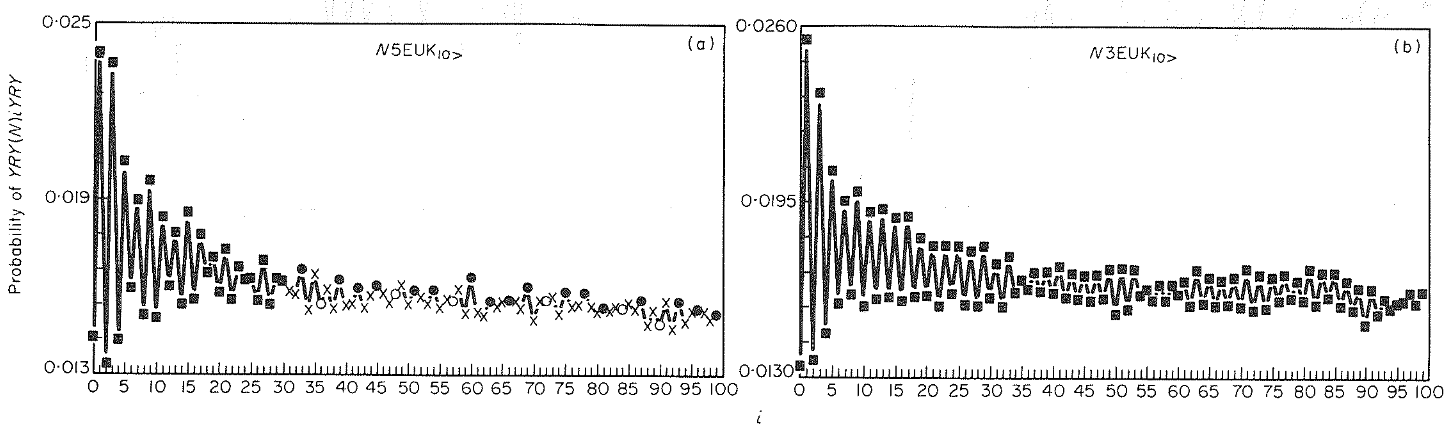


FIG. 3. Mean occurrence probability of the i-motif $YRY(N)_iYRY$ in the 5′ and 3′ eukaryotic region subpopulations which incorporate only the sequences having alternating purine/pyrimidine stretches of length > 10 bases (see method and legend in Fig. 1): (a) $N5\mathrm{EUK}_{10>}$, (b) $N3\mathrm{EUK}_{10>}$.

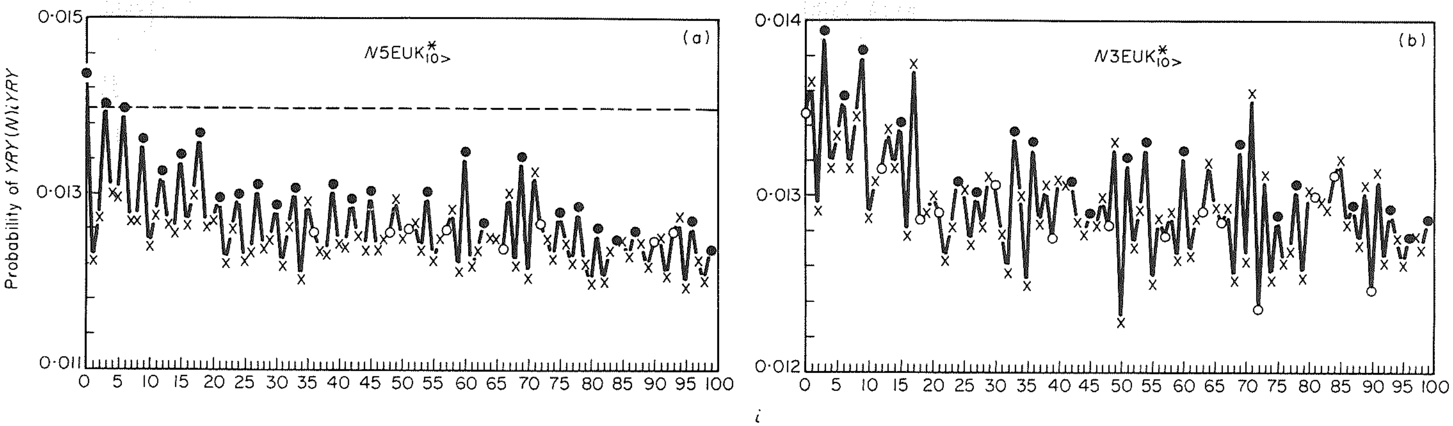


FIG. 4. Mean occurrence probability of the i-motif $YRY(N)_i YRY$ in the 5′ and 3′ eukaryotic region subpopulations $N5EUK_{10>}$ and $N3EUK_{10>}$ (see legend in Fig. 3) in which the alternating purine/pyrimidine stretches of length >10 bases are deleted (see method and legend in Fig. 1): (a) $N5EUK^*_{10>}$, (b) $N3EUK^*_{10>}$.

intensity of P2 is weak in 5′ regions (P2 in the range [0, 23] in N5EUK, P2 in the range [0, 30] in $N5\mathrm{EUK}_{10>}$) and high in 3′ regions (P2 in the range [0, 99] in N3EUK and in $N3\mathrm{EUK}_{10>}$). The intensity of P2 is medium in eukaryotic introns (P2 in the range [0, 50] in IEUK).

4. Discussion

Statistical studies of gene populations have identified the periodicity P3 (Shepherd, 1981; Fickett, 1982; Arquès & Michel, 1987*a,b,c*) in protein coding genes of any taxonomic group: eukaryotes, prokaryotes, viruses, chloroplasts, mitochondria and plasmids. This periodicity P3 was attributed to the preferential use of the *RNY* codon (Eigen, 1971; Eigen & Schuster, 1978). Indeed, based on biological properties, the *RNY* codon model was introduced to propose a primary structure for primordial protein coding genes compatible with a simple translation apparatus (Eigen, 1971; Eigen & Schuster, 1978). This periodicity P3 is found, not only in protein coding genes, but also in some introns (Arquès & Michel, 1987*c*) of viruses (*I*VIR) and mitochondria. These two intron populations have the genetic information necessary to code for proteins (Arquès & Michel, 1987*c*). Indeed, viruses use overlapping genes, both DNA strands and alternative patterns of RNA splicing in order to maximize the functions of a genome of small size (Ziff, 1980). On the other hand, many mitochondrial introns encode splicing proteins (maturases) (Lazowska *et al.*, 1980). The 5′ and 3′ regions of eukaryotes (*N*5EUK and *N*3EUK), prokaryotes (*N*5PRO and *N*3PRO) and viruses (*N*5VIR and *N*3VIR) are newly analysed gene populations having the periodicity P3 (see results). In conclusion, *the periodicity P3, which has been thought to be specific for protein coding genes, also exists in some noncoding genes.*

A different type of periodicity, i.e. the periodicity P2 (Arquès & Michel, 1987*c*), was identified in eukaryotic introns (*I*EUK). This periodicity P2 is not related to the protein coding function, but rather to regulatory functions (Arquès & Michel, 1987*c*). This periodicity P2 also exists in the newly analysed gene populations of 5′ and 3′ eukaryotic regions (*N*5EUK and *N*3EUK; see results). In conclusion, the periodicity P2 is not specific for eukaryotic introns, but according to the current state of statistical analyses, it seems to be found only in the eukaryotic genome—in agreement with an advanced function such as regulation.

The $YRY(N)_6YRY$ preferential occurrence [i.e. $p_6(F)$ has the highest value in the range [0, 99] with most of gene populations and in a few populations, $p_6(F)$ has the second or the third highest value] is found in following gene populations (because all the results are not found in this ref.):

—protein coding genes of eukaryotes, prokaryotes, viruses, chloroplasts, mitochondria and plasmids,

—introns of viruses and chloroplasts,

—ribosomal, transfer and small nuclear RNA genes.

The 5′ and 3′ prokaryotic regions (*N*5PRO and *N*3PRO) are two newly analysed gene populations having the $YRY(N)_6YRY$ preferential occurrence [see Fig. 1(b) and (g)]. The populations with the periodicity P2 (*N*5EUK, *I*EUK and *N*3EUK) cannot have the $YRY(N)_6YRY$ preferential occurrence. Nevertheless, this problem remains open because, in eukaryotic introns, the $YRY(N)_6YRY$ preferential occurrence is hidden by the periodicity P2 (Arquès & Michel, 1987*c*). This situation is also observed in the 5′ eukaryotic regions with $N5EUK^*$ [see Fig. 2(a)] and $N5EUK^*_{10>}$ [see Fig. 4(a)].

Finally, these observations suggest a unitary and dynamic concept for the genes because for a given genome, the 5′ and 3′ regions have the genetic information for protein coding genes and for introns (see Table 3):

(1) In the eukaryotic genome, the 5′ (P2 and P3) and 3′ (P2 and P3) regions have the information for protein coding genes (P3) and for introns (P2). According to the intensities of P2 and P3 (see Results), P2 seems to move in the 3′-5′ direction, while P3, in the 5′-3′ direction.

(2) In the prokaryotic genome, the 5′ (P3) and 3′ (P3) regions have the information for protein coding genes (P3).

(3) In the viral genome, the 5′ (P3) and 3′ (P3) regions have the information for protein coding genes (P3) and for introns (P3). The absence of P2 in the viral introns (in opposition to eukaryotic introns) may be related to the absence of P2 in 5′ and 3′ regions of viruses.

TABLE 3

Periodicities P2 and P3 in gene populations

	5′ regions	Coding genes	Introns	3′ regions
Eukaryotes	P2(+) and P3(++)	P3(+++)	P2(++)	P2(+++) and P3(+)
Prokaryotes	P3	P3	—	P3
Viruses	P3	P3	P3	P3

(The symbols "+" indicate the intensity of the periodicities in the eukaryotic genome.)

We thank Professors Max Burger, Herman Moser and Jacques Streith, Dr Christoph Nager, Thomas Nyffenegger and Nouchine Soltanifar for their advice, and the bioinformatic group for its assistance. This work was supported by grants from the Unité Associée CNRS No. 822 to D.G.A. and from the Friedrich Miescher Institute to C.J.M.

REFERENCES

ARQUÈS, D. G. & MICHEL, C. J. (1987*a*). *Math. Biosc.* **86**, 1.
ARQUÈS, D. G. & MICHEL, C. J. (1987*b*). *J. theor. Biol.* **128**, 457.
ARQUÈS, D. G. & MICHEL, C. J. (1987*c*). *Nucl. Acids Res.* **15**, 7581.
EIGEN, M. (1971). *Naturwissenschaften* **58**, 465.
EIGEN, M. & SCHUSTER, P. (1978). *Naturwissenschaften* **65**, 341.
FICKETT, J. W. (1982). *Nucl. Acids Res.* **10**, 5303.
LAZOWSKA, J., JACQ, C. & SLONIMSKI, P. P. (1980). *Cell* **22**, 333.
SHEPHERD, J. C. W. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78**, 1596.
ZIFF, E. B. (1980). *Nature, Lond.* **287**, 491.